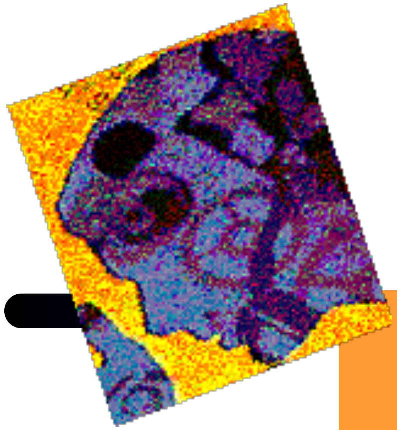
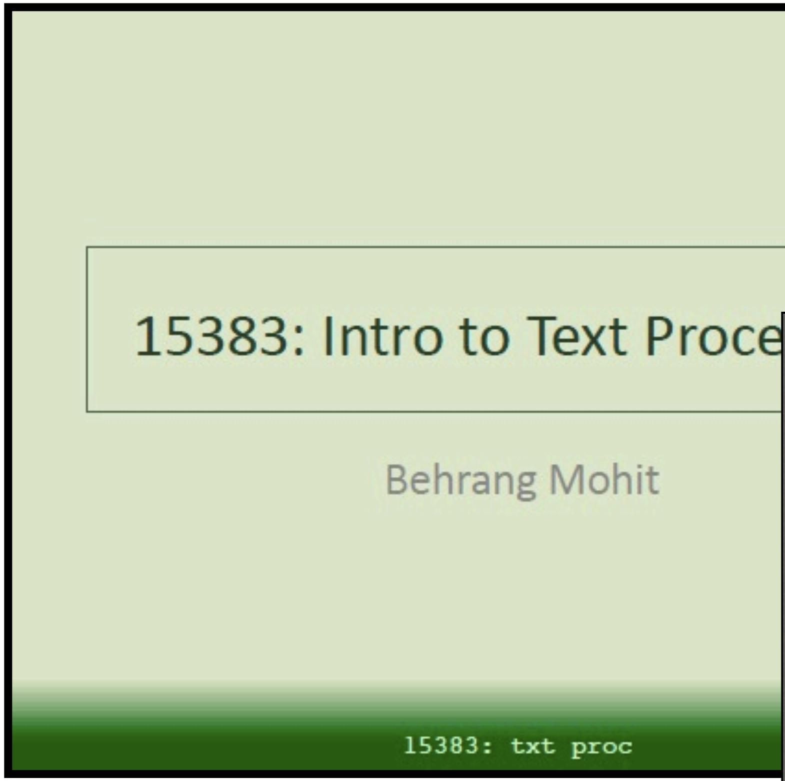


# *Natural Language Processing Computational Linguistics Text processing*

# *Content*

- Acknowledgments
- Examples
- Defintions
- History
- Objective
- Levels - Problems
- Applications

# Acknowledgment



15383: Intro to Text Proce

Behrang Mohit

15383: txt proc

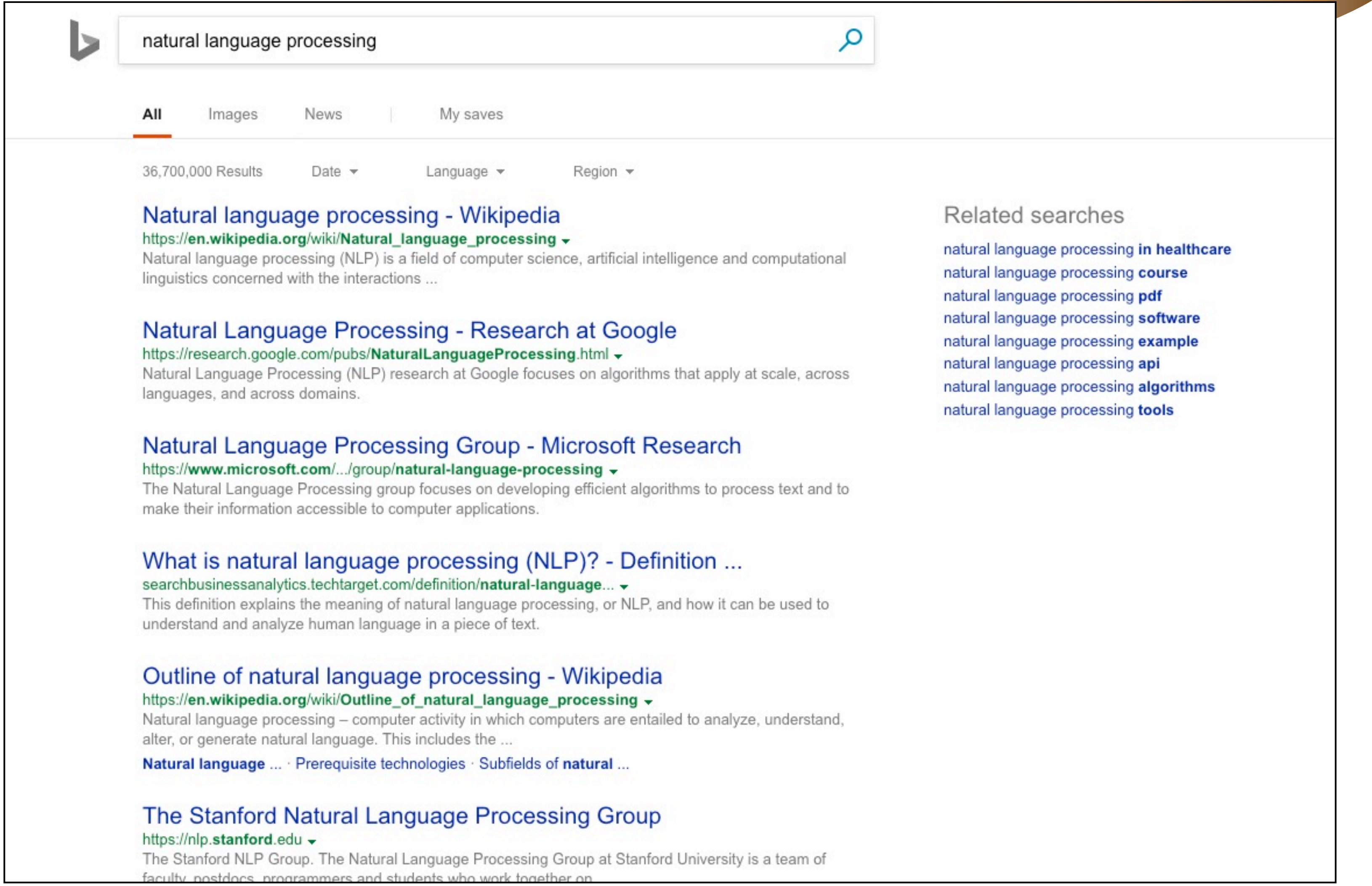
Natural Language Processing (**NLP**)

Traitement automatique des langues naturelles (**TALN**)

المعالجة الآلية للغات الطبيعية

Azzeddine Mazroui
Master d'Ingénierie Informatique
M2I
2016-2017

# Examples



natural language processing

36,700,000 Results

Natural language processing - Wikipedia
https://en.wikipedia.org/wiki/Natural_language_processing
Natural language processing (NLP) is a field of computer science, artificial intelligence and computational linguistics concerned with the interactions ...

Natural Language Processing - Research at Google
https://research.google.com/pubs/NaturalLanguageProcessing.html
Natural Language Processing (NLP) research at Google focuses on algorithms that apply at scale, across languages, and across domains.

Natural Language Processing Group - Microsoft Research
https://www.microsoft.com/.../group/natural-language-processing
The Natural Language Processing group focuses on developing efficient algorithms to process text and to make their information accessible to computer applications.

What is natural language processing (NLP)? - Definition ...
searchbusinessanalytics.techtarget.com/definition/natural-language...
This definition explains the meaning of natural language processing, or NLP, and how it can be used to understand and analyze human language in a piece of text.

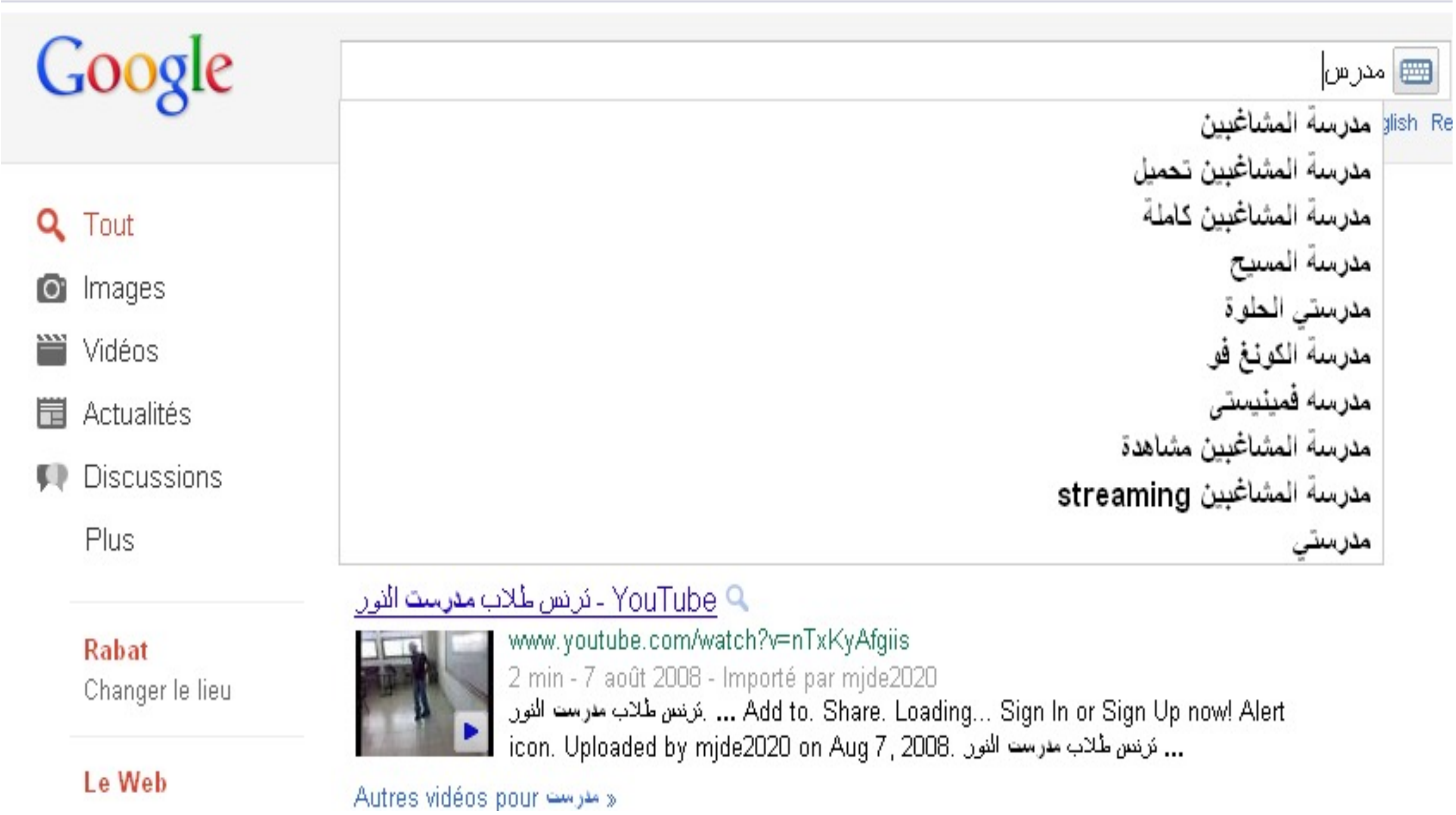
Outline of natural language processing - Wikipedia
https://en.wikipedia.org/wiki/Outline_of_natural_language_processing
Natural language processing – computer activity in which computers are entailed to analyze, understand, alter, or generate natural language. This includes the ...
Natural language ... · Prerequisite technologies · Subfields of natural ...

The Stanford Natural Language Processing Group
https://nlp.stanford.edu
The Stanford NLP Group. The Natural Language Processing Group at Stanford University is a team of faculty, postdocs, programmers and students who work together on

Related searches

- natural language processing **in healthcare**
- natural language processing **course**
- natural language processing **pdf**
- natural language processing **software**
- natural language processing **example**
- natural language processing **api**
- natural language processing **algorithms**
- natural language processing **tools**

# *Examples*



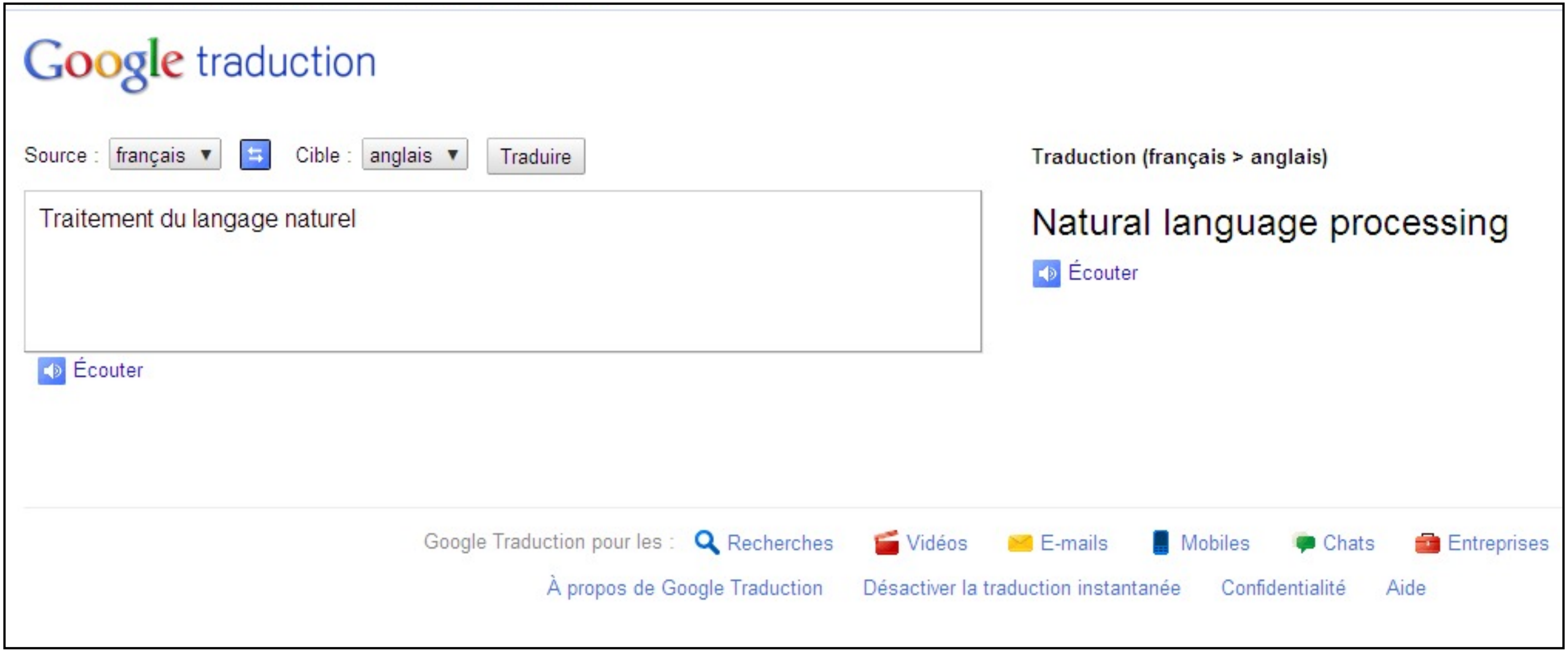
مدرس

مدرسة المشاغبين
مدرسة المشاغبين تحميل
مدرسة المشاغبين كاملة
مدرسة المسيح
مدرستي الحلوة
مدرسة الكونغ فو
مدرسة فمينيستى
مدرسة المشاغبين مشاهدة
مدرسة المشاغبين streaming
مدرستي

ترنس طلاب مدرست النور - YouTube
www.youtube.com/watch?v=nTxKyAfgiis
2 min - 7 août 2008 - Importé par mjde2020
ترنس طلاب مدرست النور. ... Add to. Share. Loading... Sign In or Sign Up now! Alert icon. Uploaded by mjde2020 on Aug 7, 2008. ترنس طلاب مدرست النور ...

Autres vidéos pour مدرست »

# *Examples*



Google traduction

Source : français ▼ ⇆ Cible : anglais ▼ Traduire

Traitement du langage naturel

Écouter

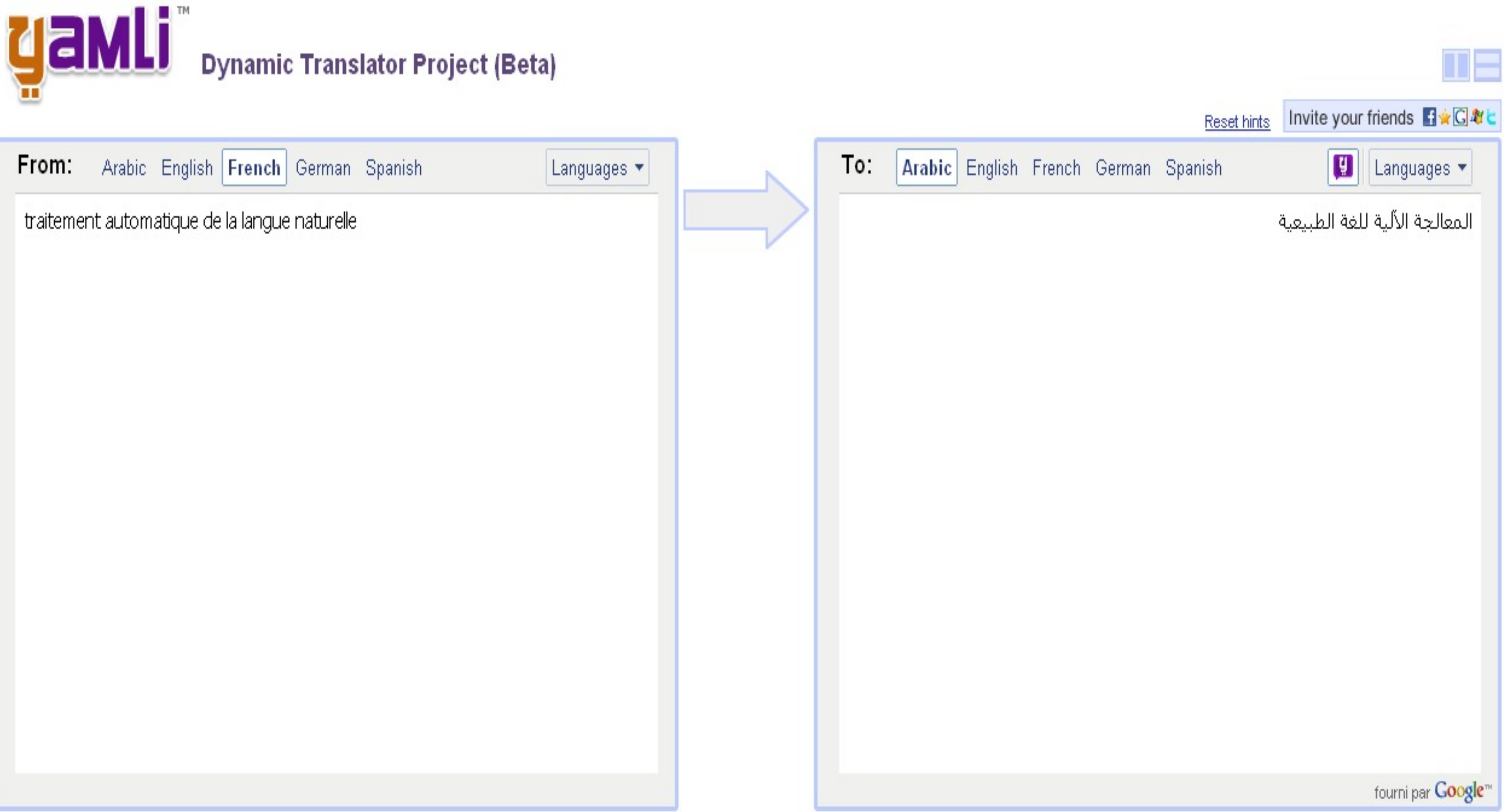
Traduction (français > anglais)

Natural language processing

Écouter

Google Traduction pour les : Recherches Vidéos E-mails Mobiles Chats Entreprises
À propos de Google Traduction
Désactiver la traduction instantanée
Confidentialité
Aide

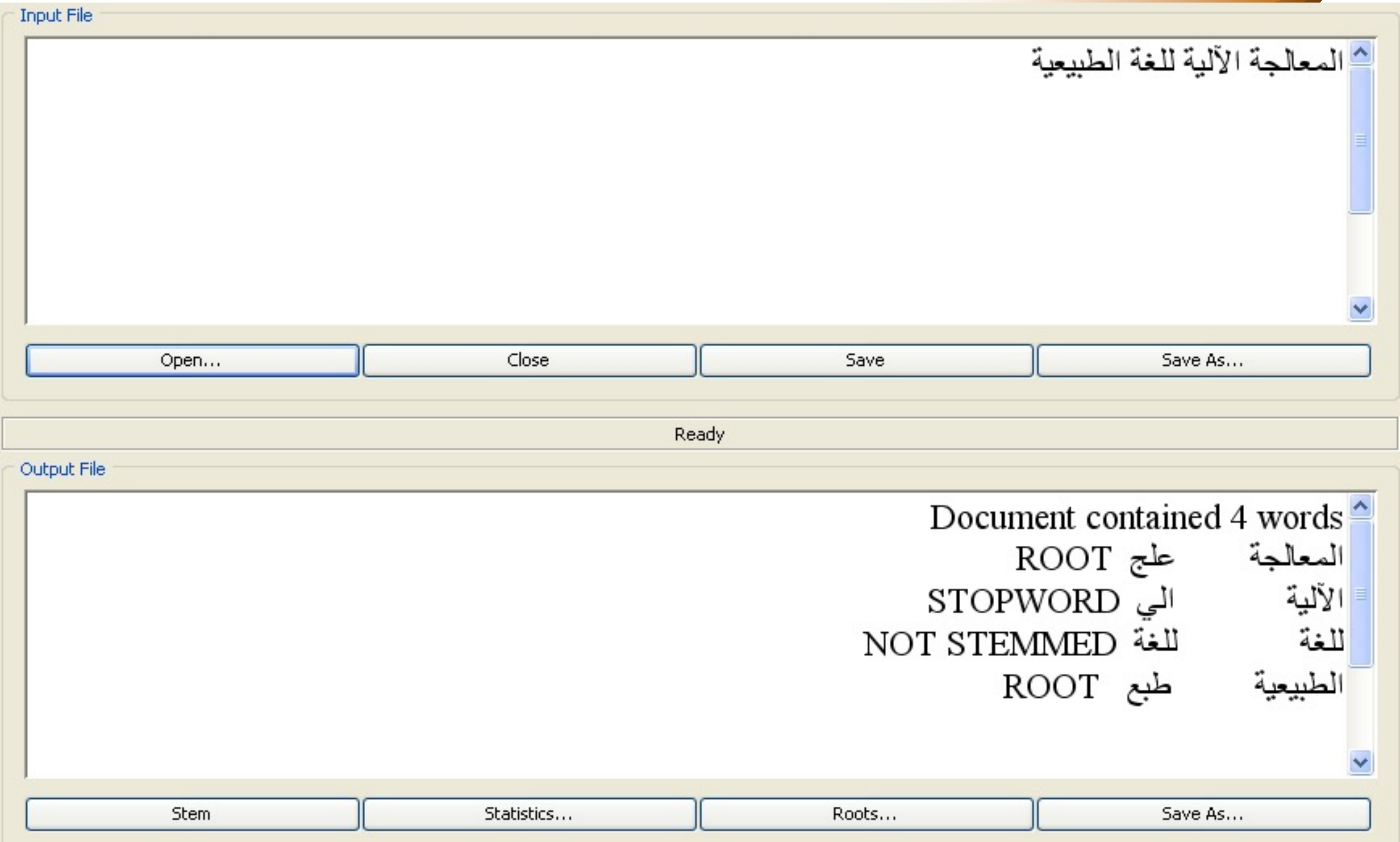
# *Examples*


yamli™
Dynamic Translator Project (Beta)
Reset hints
Invite your friends
From: Arabic English French German Spanish Languages ▾
To: Arabic English French German Spanish Languages ▾
fourni par Google™

traitement automatique de la langue naturelle

المعالجة الآلية للغة الطبيعية

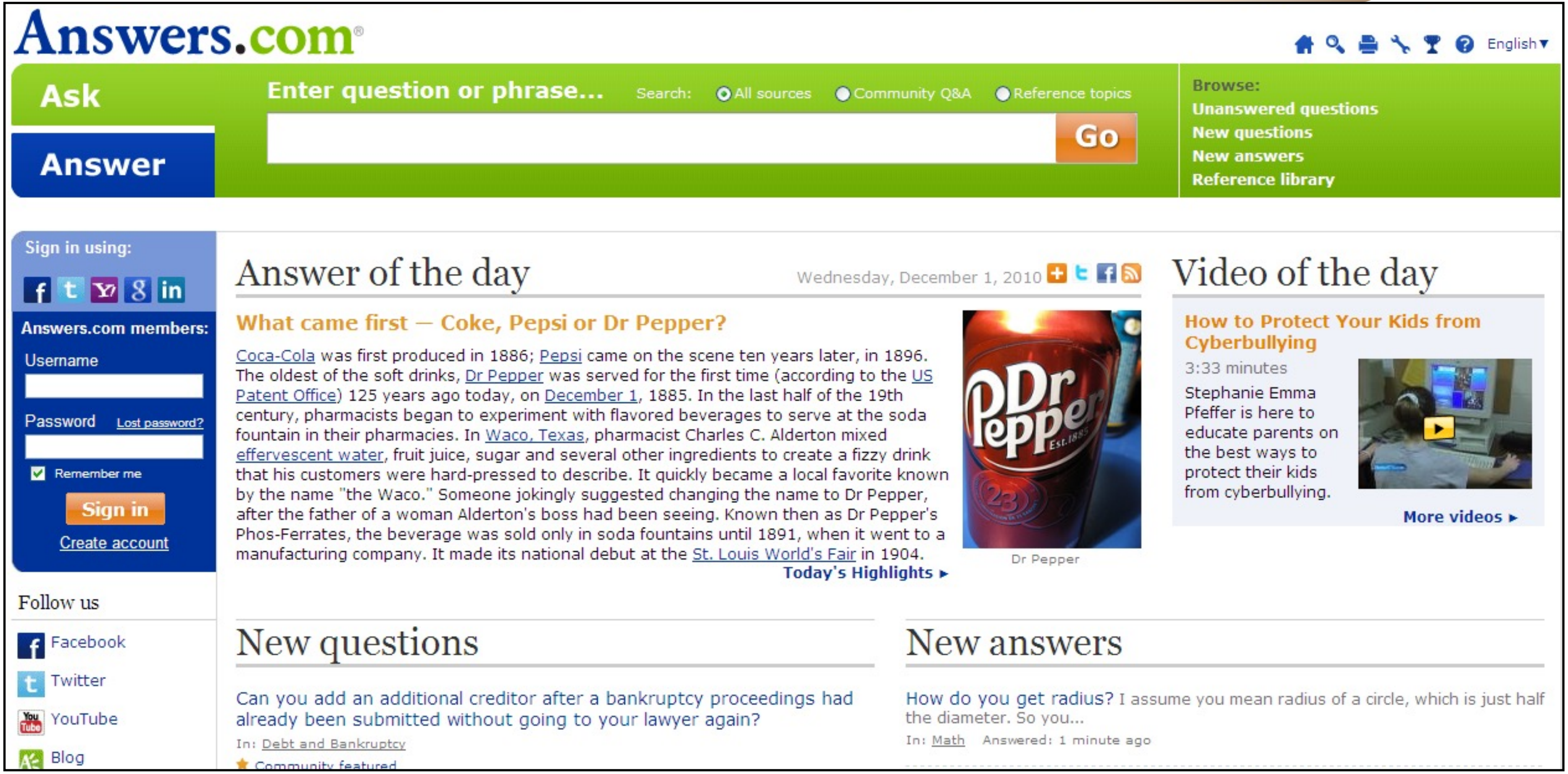
# Examples



المعالجة الآلية للغة الطبيعية

```
Document contained 4 words
المعالجة      علج ROOT
الآلية      الي STOPWORD
للغة      للغة NOT STEMMED
الطبيعية      طبع ROOT
```

# *Examples*



## Answer of the day

Wednesday, December 1, 2010

### What came first — Coke, Pepsi or Dr Pepper?

Coca-Cola was first produced in 1886; Pepsi came on the scene ten years later, in 1896. The oldest of the soft drinks, Dr Pepper was served for the first time (according to the US Patent Office) 125 years ago today, on December 1, 1885. In the last half of the 19th century, pharmacists began to experiment with flavored beverages to serve at the soda fountain in their pharmacies. In Waco, Texas, pharmacist Charles C. Alderton mixed effervescent water, fruit juice, sugar and several other ingredients to create a fizzy drink that his customers were hard-pressed to describe. It quickly became a local favorite known by the name "the Waco." Someone jokingly suggested changing the name to Dr Pepper, after the father of a woman Alderton's boss had been seeing. Known then as Dr Pepper's Phos-Ferrates, the beverage was sold only in soda fountains until 1891, when it went to a manufacturing company. It made its national debut at the St. Louis World's Fair in 1904.

Today's Highlights ▸

Dr Pepper

## Video of the day

### How to Protect Your Kids from Cyberbullying

3:33 minutes

Stephanie Emma Pfeffer is here to educate parents on the best ways to protect their kids from cyberbullying.

More videos ▸

## New questions

Can you add an additional creditor after a bankruptcy proceedings had already been submitted without going to your lawyer again?

In: Debt and Bankruptcy

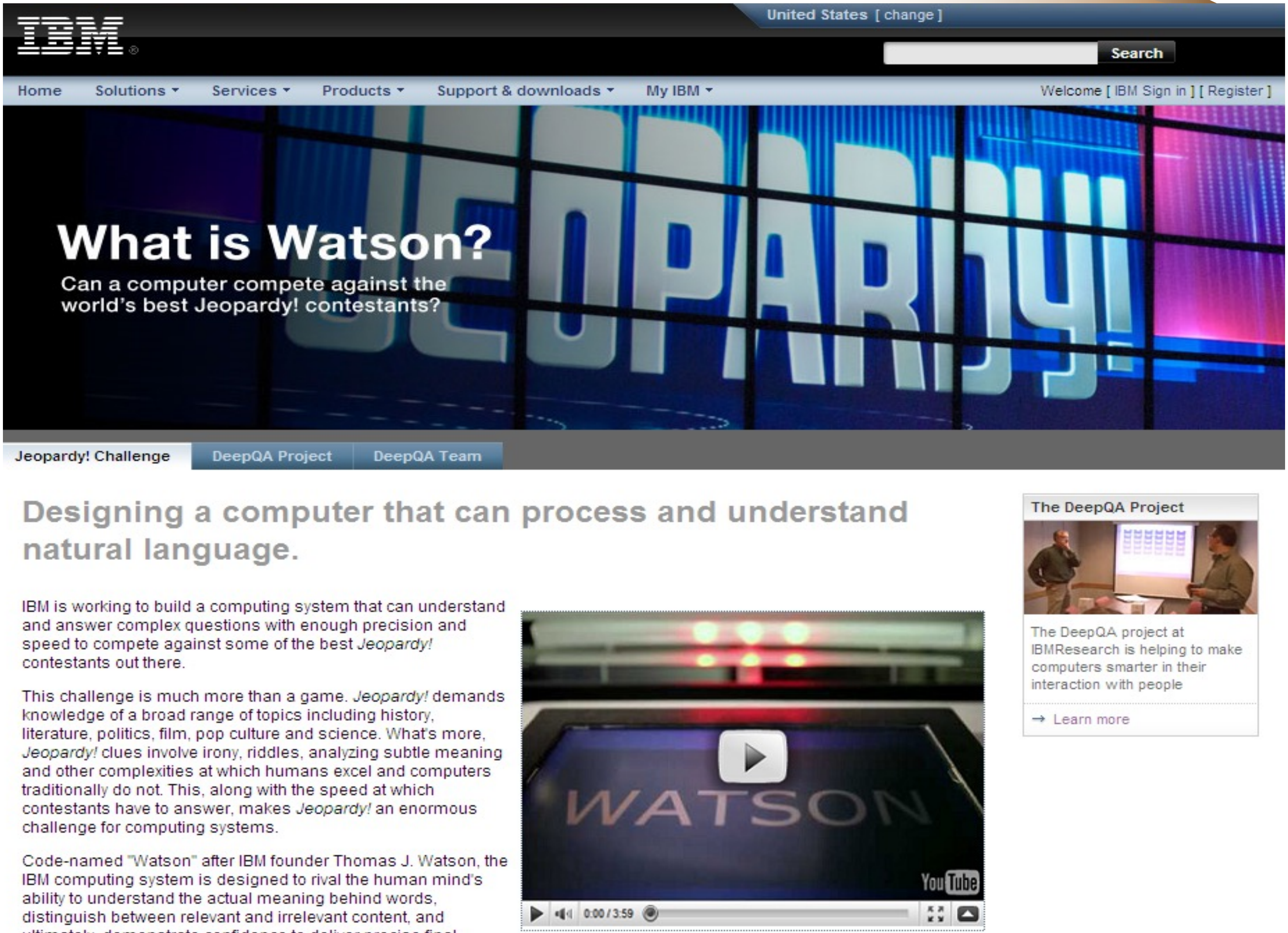
★ Community featured

## New answers

How do you get radius? I assume you mean radius of a circle, which is just half the diameter. So you...

In: Math Answered: 1 minute ago

# Examples



What is Watson?

Can a computer compete against the world's best Jeopardy! contestants?

## Designing a computer that can process and understand natural language.

IBM is working to build a computing system that can understand and answer complex questions with enough precision and speed to compete against some of the best *Jeopardy!* contestants out there.

This challenge is much more than a game. *Jeopardy!* demands knowledge of a broad range of topics including history, literature, politics, film, pop culture and science. What's more, *Jeopardy!* clues involve irony, riddles, analyzing subtle meaning and other complexities at which humans excel and computers traditionally do not. This, along with the speed at which contestants have to answer, makes *Jeopardy!* an enormous challenge for computing systems.

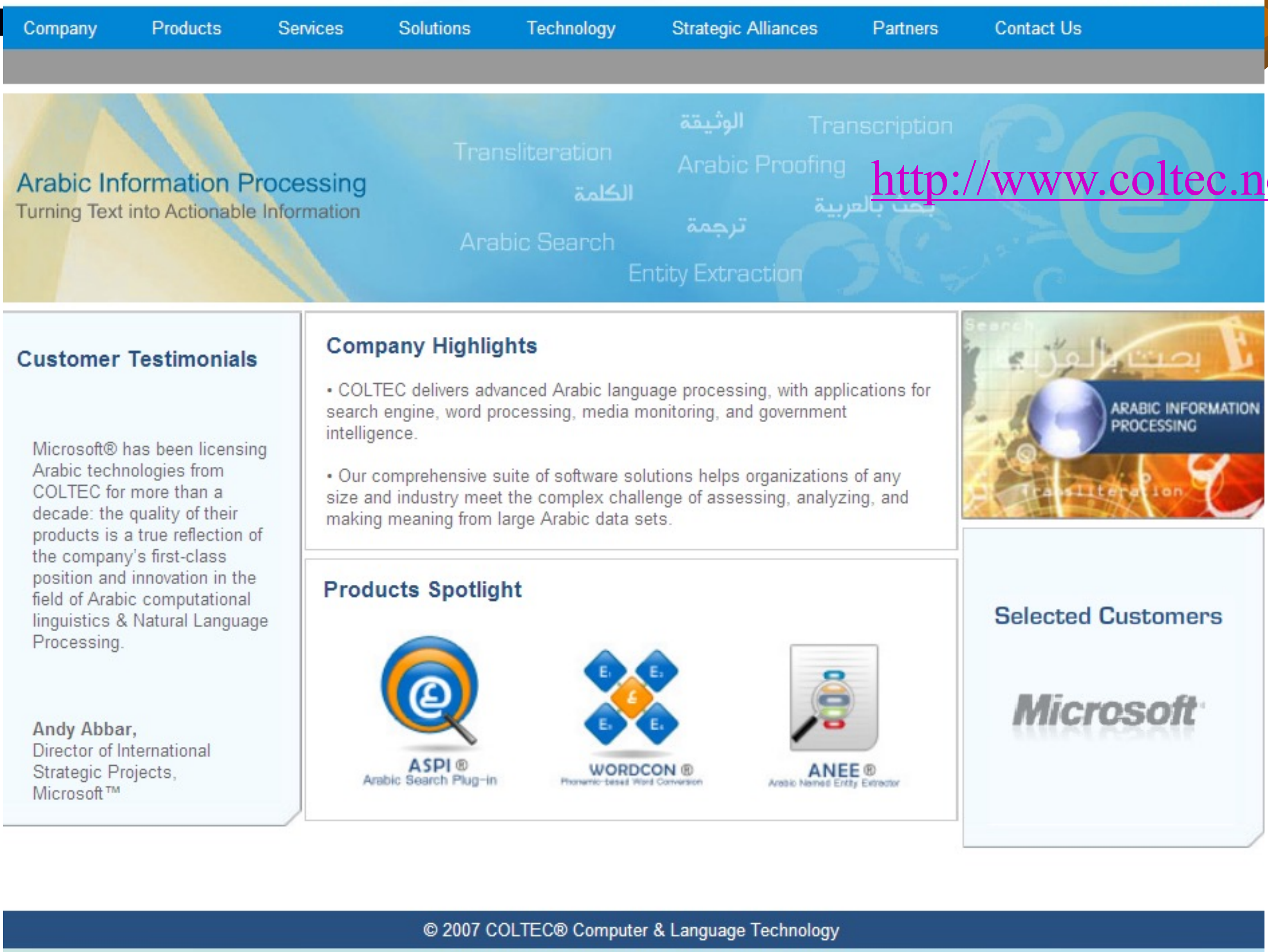
Code-named "Watson" after IBM founder Thomas J. Watson, the IBM computing system is designed to rival the human mind's ability to understand the actual meaning behind words, distinguish between relevant and irrelevant content, and ultimately, demonstrate confidence to deliver precise final

The DeepQA Project

The DeepQA project at IBMResearch is helping to make computers smarter in their interaction with people

→ Learn more

# *Examples*



Company | Products | Services | Solutions | Technology | Strategic Alliances | Partners | Contact Us

## Arabic Information Processing

Turning Text into Actionable Information

[http://www.coltec.net/](http://www.coltec.net/)

### Customer Testimonials

Microsoft® has been licensing Arabic technologies from COLTEC for more than a decade: the quality of their products is a true reflection of the company's first-class position and innovation in the field of Arabic computational linguistics & Natural Language Processing.

**Andy Abbar,**
Director of International Strategic Projects,
Microsoft™

### Company Highlights

- COLTEC delivers advanced Arabic language processing, with applications for search engine, word processing, media monitoring, and government intelligence.
- Our comprehensive suite of software solutions helps organizations of any size and industry meet the complex challenge of assessing, analyzing, and making meaning from large Arabic data sets.

### Products Spotlight

### Selected Customers

© 2007 COLTEC® Computer & Language Technology

# *Defintion*

The human does not have a stock of possible sentences but a set of rules and principles that make it possible to analyze and generate any sentence of the language. It is such a system that is the subject of linguistic studies and computational linguistics

# *Defintion*

The term natural language processing (NLP) refers to all research and development aimed at modeling and reproducing, using machines, the human capacity to produce and understand linguistic utterances for communication purposes

# *Defintion*

NLP implements tools and techniques that fall under:

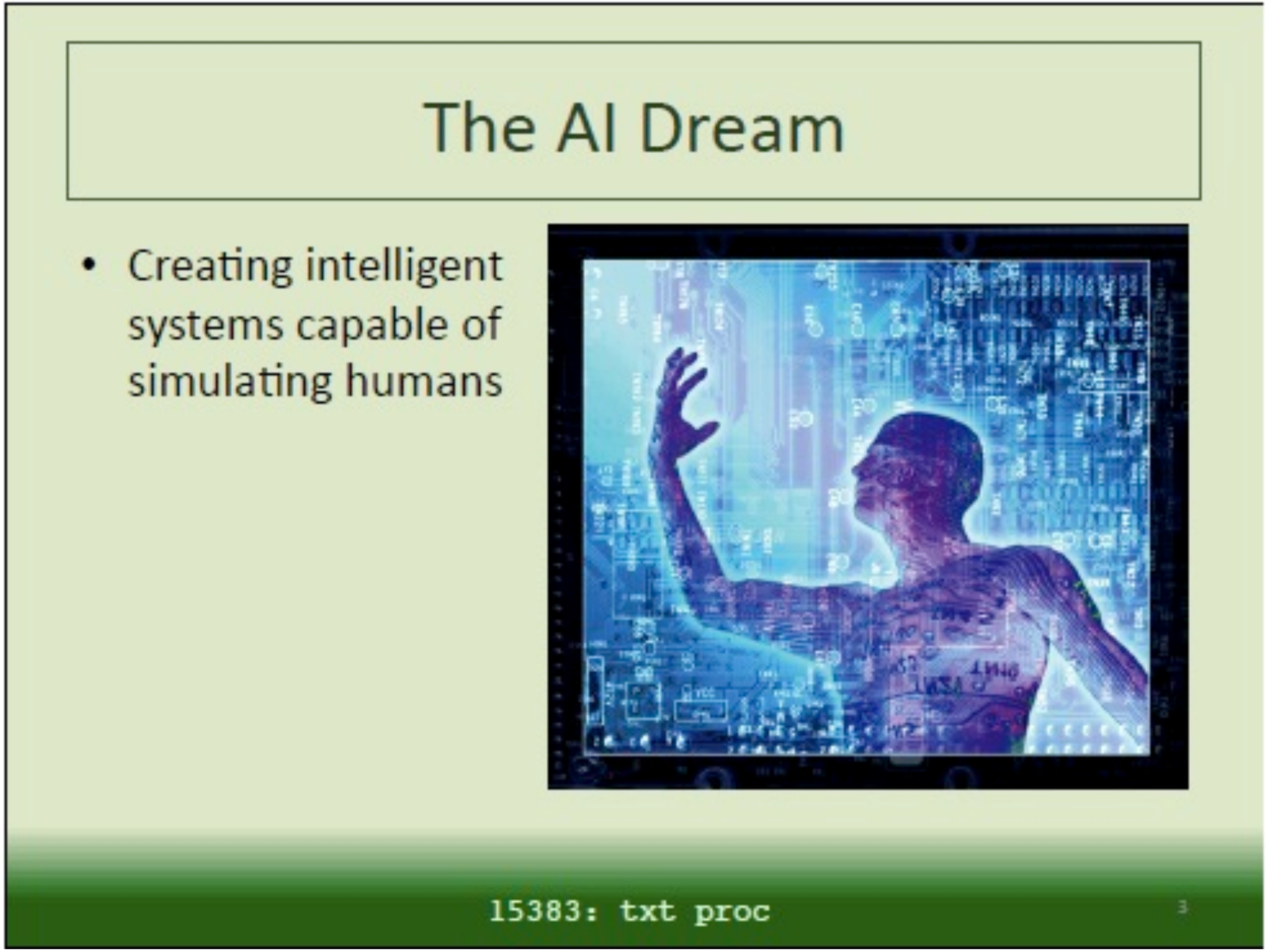
- linguistics (provide fully explicit descriptions)
- computer science (to optimize algorithms and programs)
- mathematics: algebra, logic, statistics, ... (define formal properties of processing tools and linguistic theories)
- artificial intelligence, experimental psychology, (representing knowledge)

# History of AI

- 1943 McCulloch & Pitts: Boolean circuit model of brain
- 1950 Turing's "Computing Machinery and Intelligence“
- 1956 Dartmouth meeting: "Artificial Intelligence" adopted
- 1952—69 Big hopes!
  - Newell and Simon: GPS (General Problem Solver)
  - McCarty: LISP
  - Minsky: Micro-Worlds
- 1966—73 AI discovers computational complexity
  Neural network research almost disappears
  The problem is not as easy as we thought
- 1969—79 Early development of knowledge-based systems
  Expert systems
  Ed Feigenbaum (Stanford): Knowledge is power!
  - Dendral (inferring molecular structure from a mass spectrometer).
  - MYCIN: diagnosis of blood infections

  Robotic vision applications
- 1980-- AI becomes an industry
- 1986-- Neural networks return to popularity
- 1987-- AI becomes a science
- 1995-- The emergence of intelligent agents

# History



## The AI Dream

- Creating intelligent systems capable of simulating humans

15383: txt proc

3

# History

## Language and Text

- Has been present since early days of human civilization.

Doctors appt at 4pm. Call me to discuss after school ok?

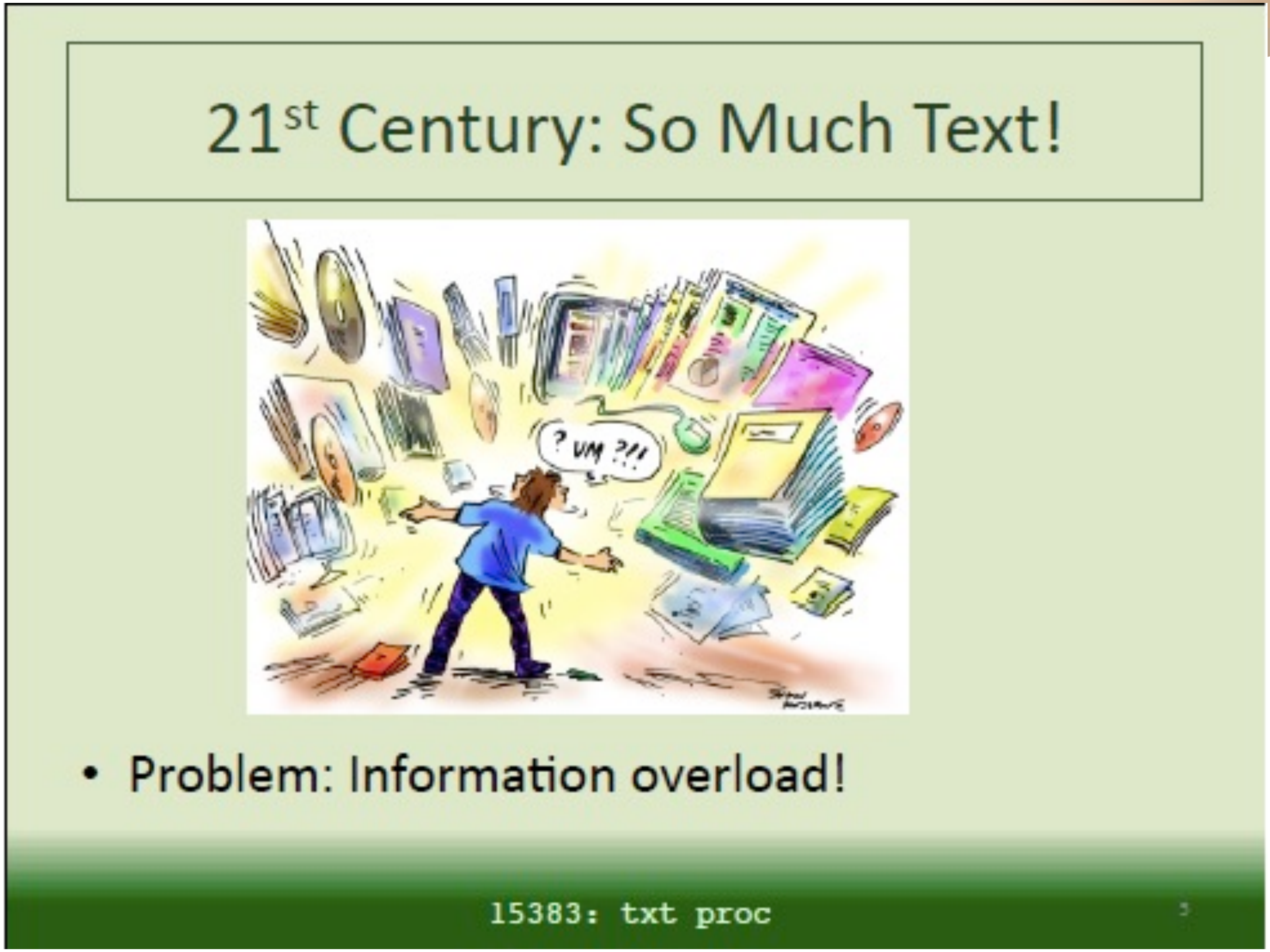
Happy Birthday!!

Maya
Messages
Call
Text Message
space
return

15383: txt proc

4

# History

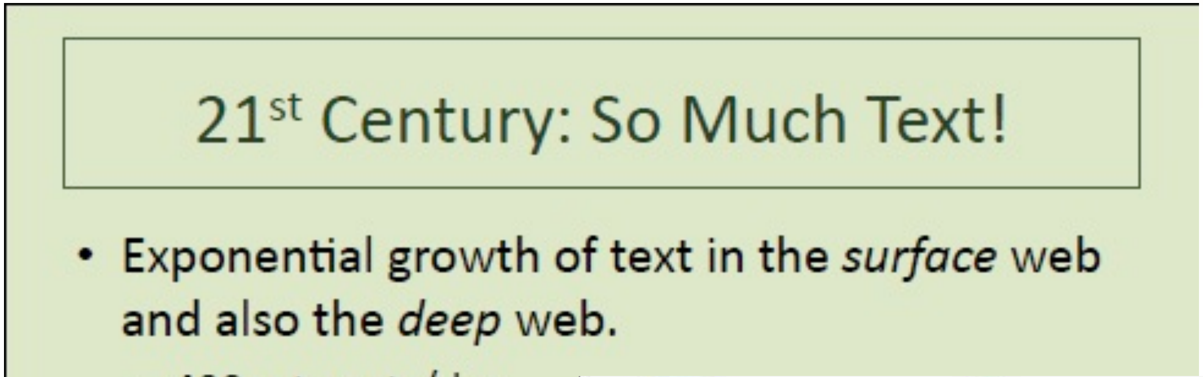


## 21st Century: So Much Text!

- Problem: Information overload!

15383: txt proc

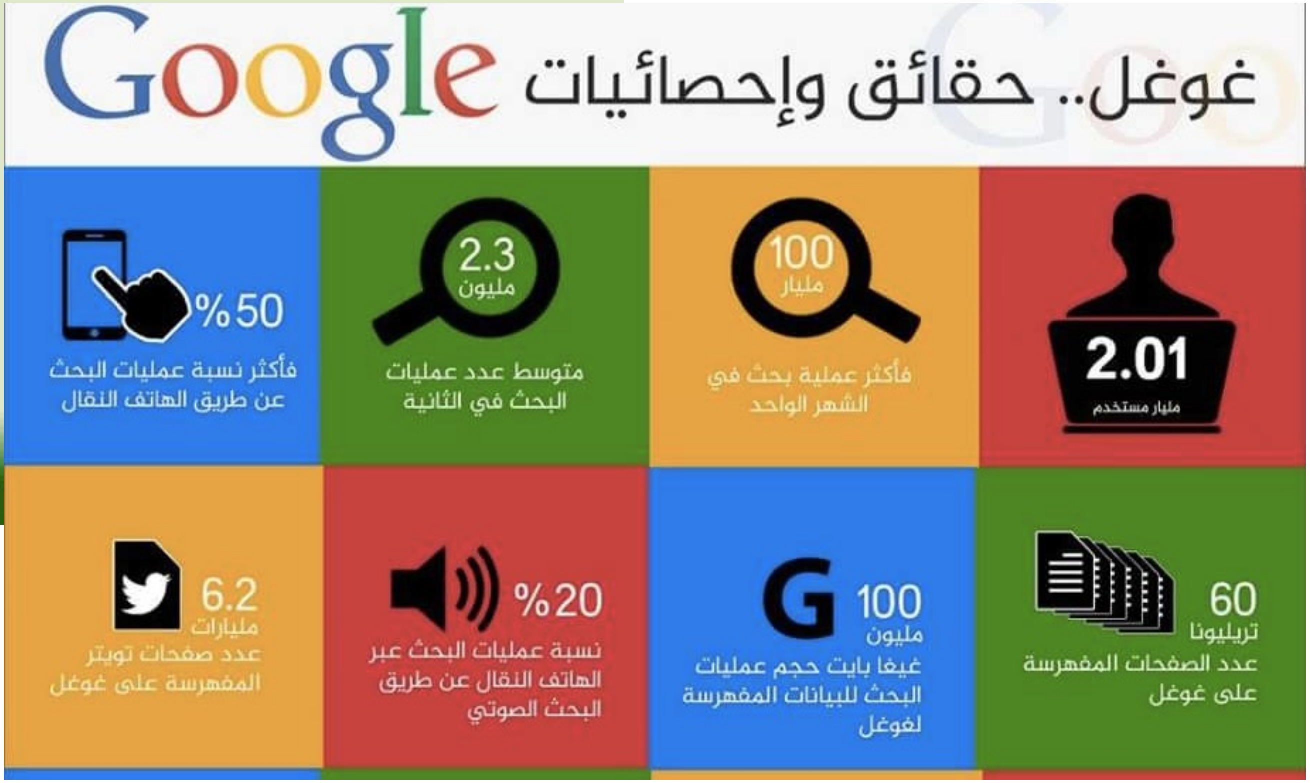
# History



## 21st Century: So Much Text!

- Exponential growth of text in the *surface* web and also the *deep* web.
  - – 400m tweets/day

15383:



غوغل.. حقائق وإحصائيات

2.01
مليار مستخدم

100
مليار
فأكثر عملية بحث في الشهر الواحد

2.3
مليون
متوسط عدد عمليات البحث في الثانية

%50
فأكثر نسبة عمليات البحث عن طريق الهاتف النقال

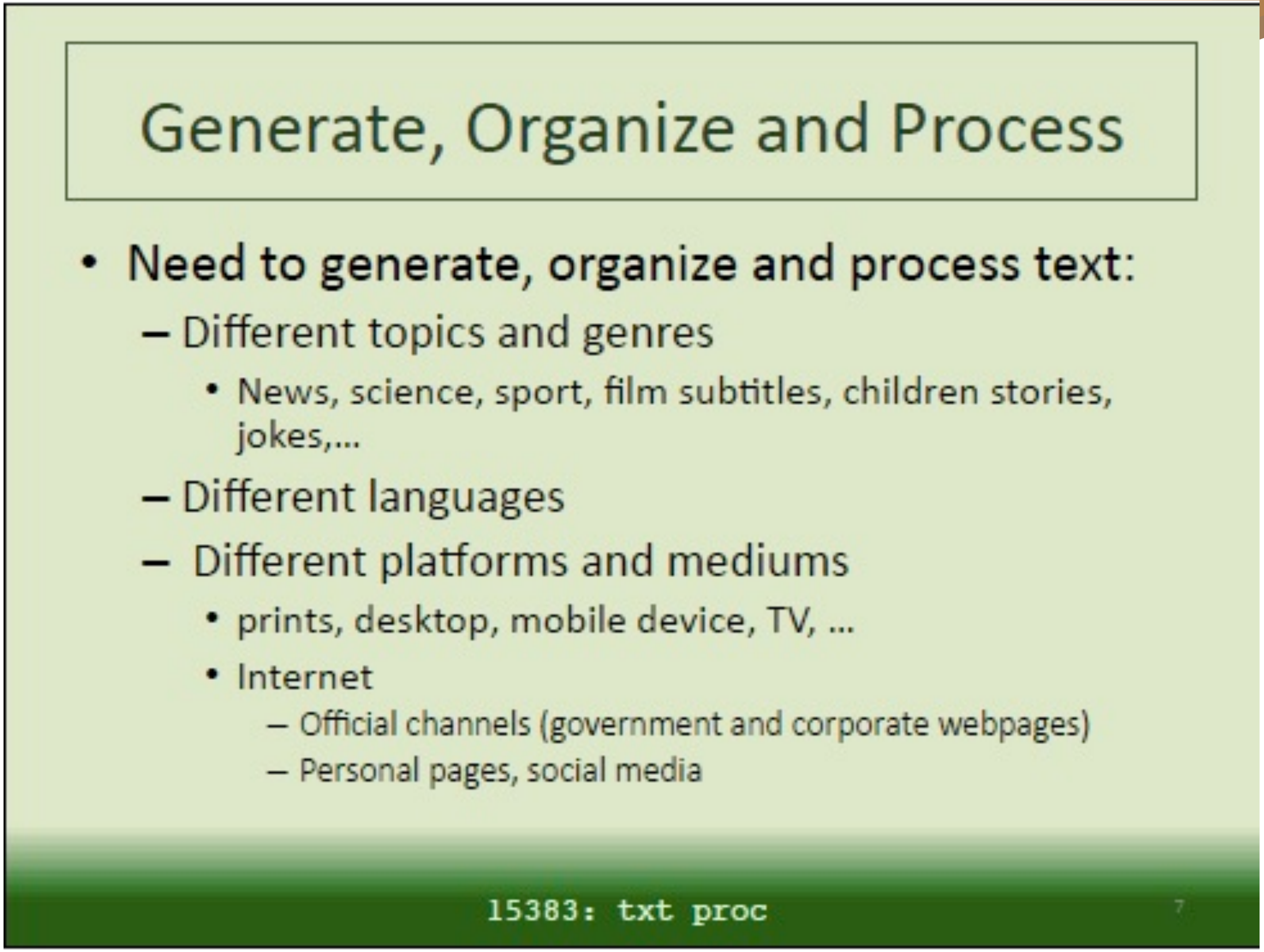
60
تريليونا
عدد الصفحات المفهرسة على غوغل

100
مليون
غيغا بايت حجم عمليات البحث للبيانات المفهرسة لغوغل

%20
نسبة عمليات البحث عبر الهاتف النقال عن طريق البحث الصوتي

6.2
مليارات
عدد صفحات تويتر المفهرسة على غوغل

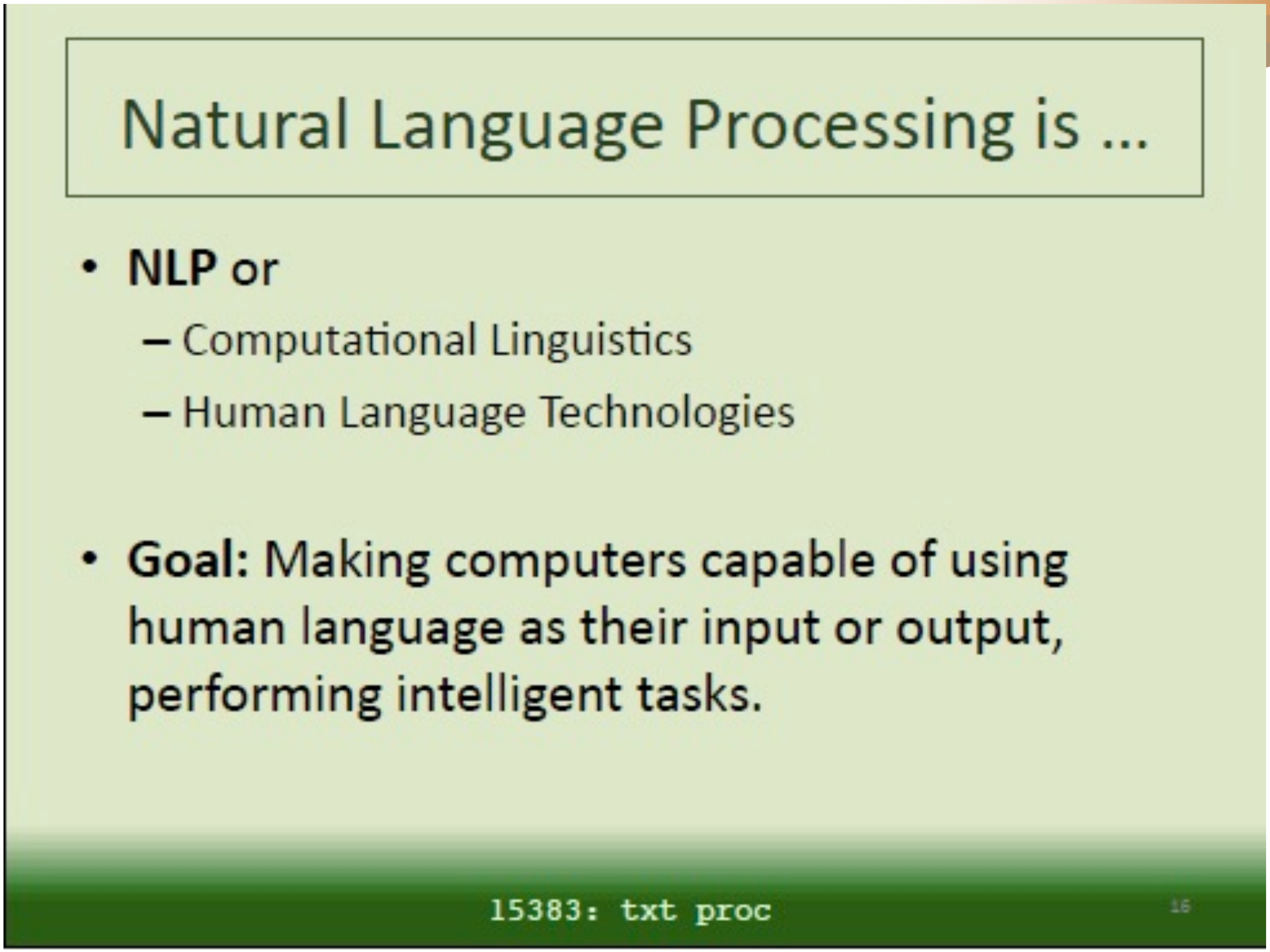
# *Objective*



## Generate, Organize and Process

- **Need to generate, organize and process text:**
  - – Different topics and genres
    - News, science, sport, film subtitles, children stories, jokes,…
  - – Different languages
  - – Different platforms and mediums
    - prints, desktop, mobile device, TV, …
    - Internet
      - – Official channels (government and corporate webpages)
      - – Personal pages, social media

# *Objective*

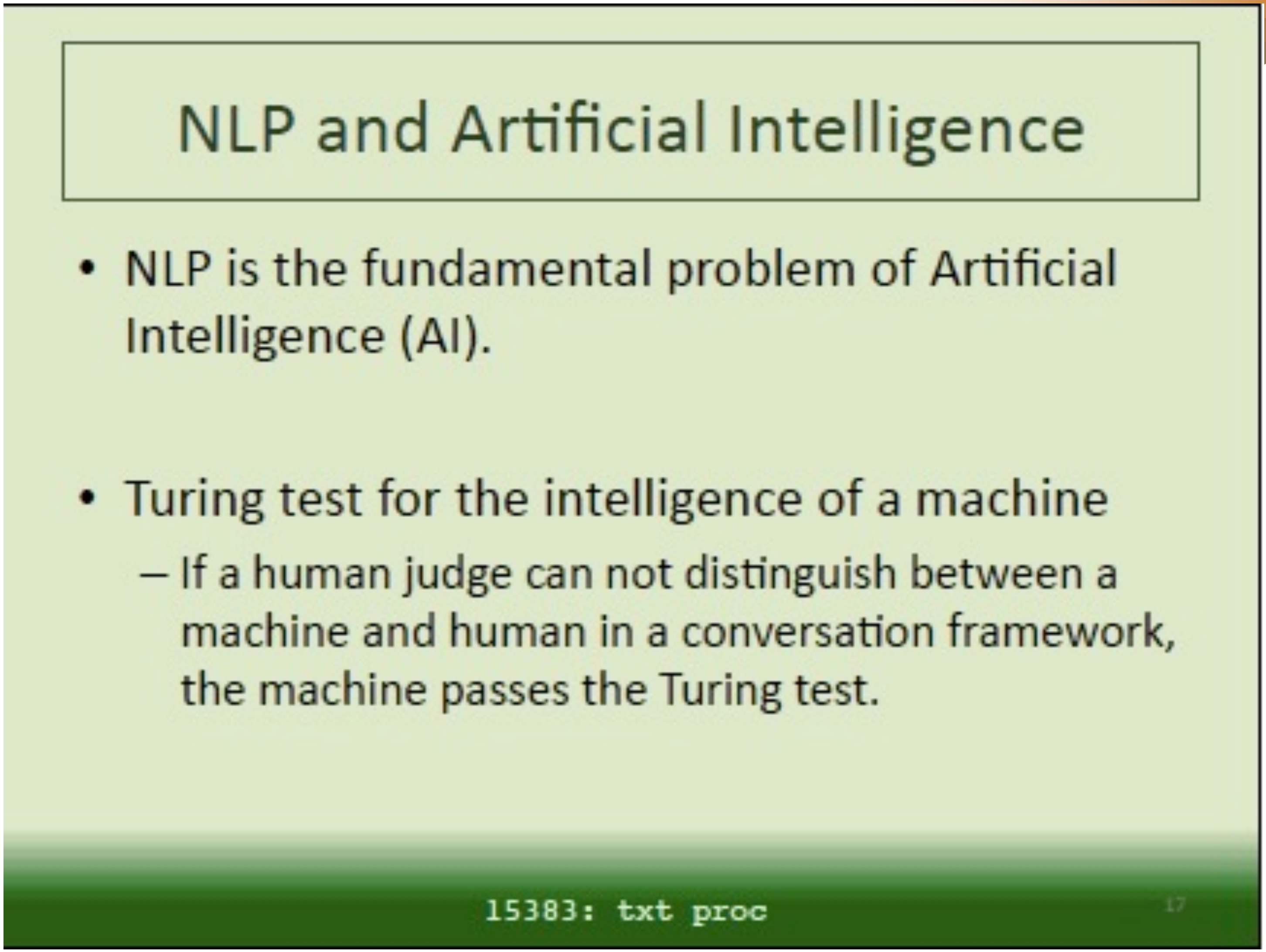


## Natural Language Processing is …

- **NLP** or
  - Computational Linguistics
  - Human Language Technologies

- **Goal:** Making computers capable of using human language as their input or output, performing intelligent tasks.

# *Objective*

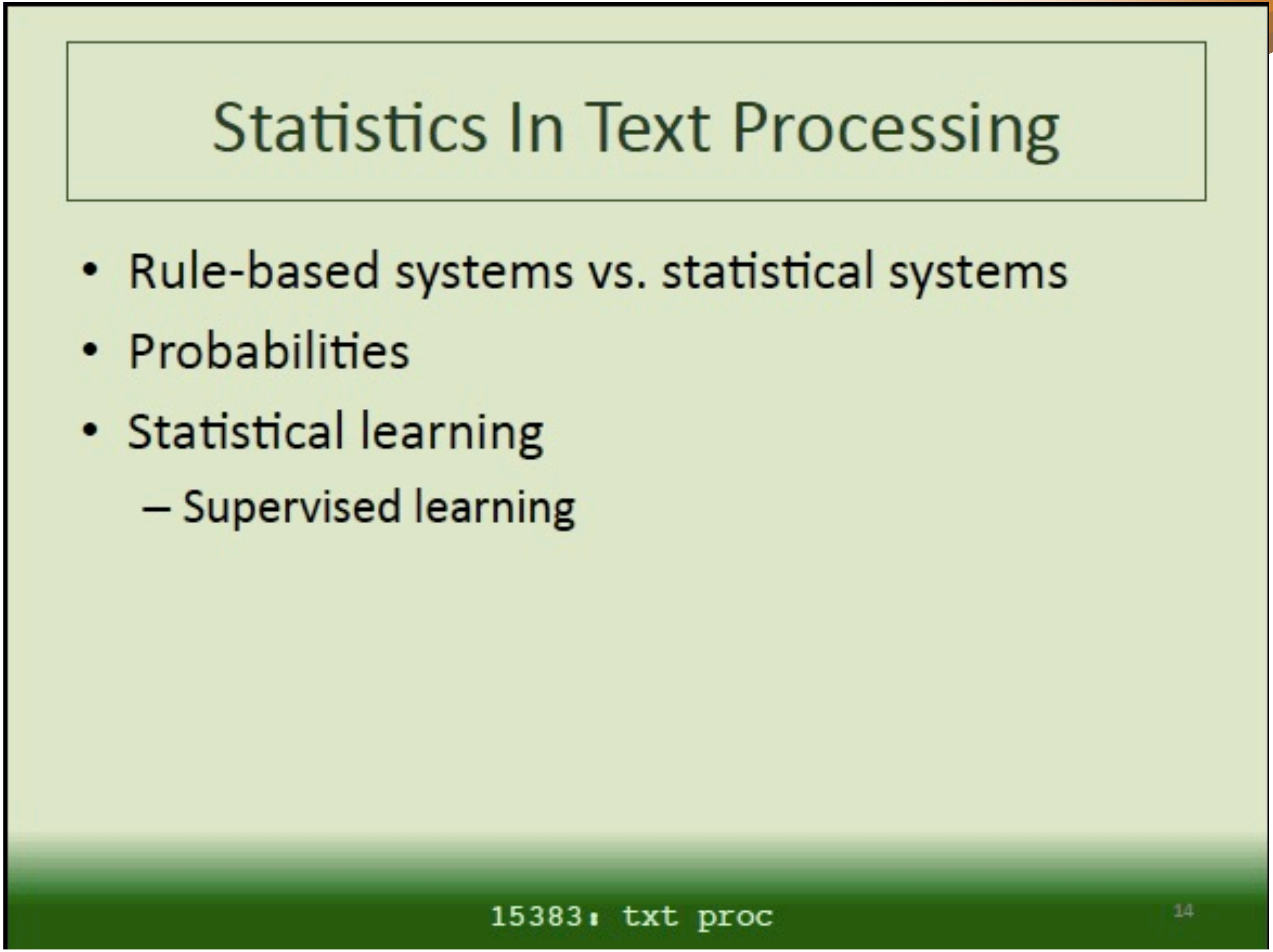


## NLP and Artificial Intelligence

- NLP is the fundamental problem of Artificial Intelligence (AI).

- Turing test for the intelligence of a machine
  - If a human judge can not distinguish between a machine and human in a conversation framework, the machine passes the Turing test.

# Content of the course

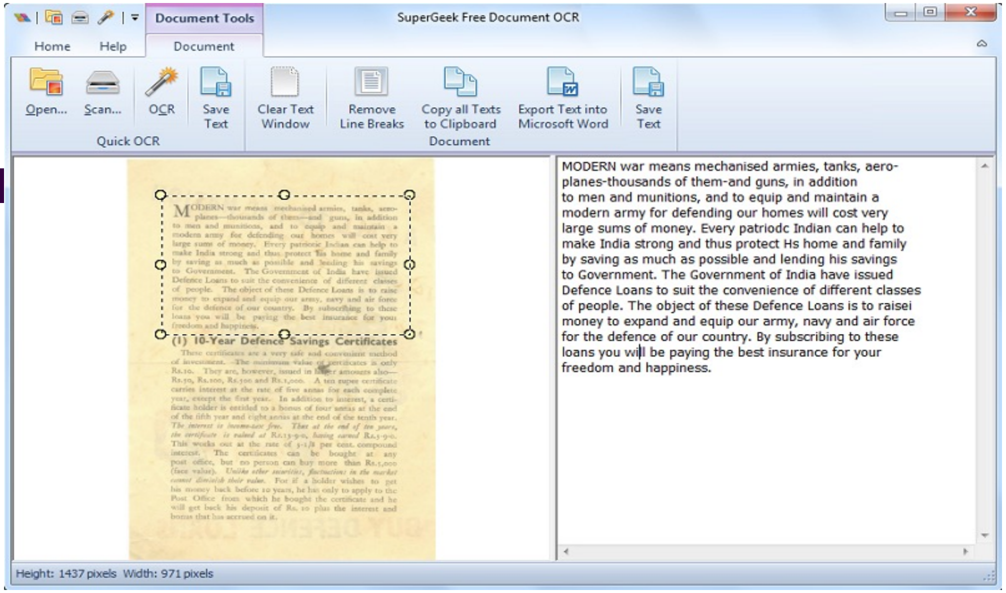


## Statistics In Text Processing

- Rule-based systems vs. statistical systems
- Probabilities
- Statistical learning
  - Supervised learning

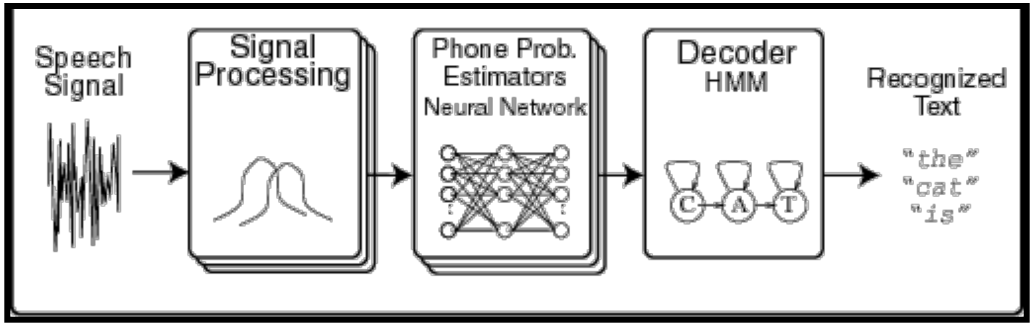
# Levels

- **Image** - OCR


Document Tools
SuperGeek Free Document OCR
Home
Help
Document
Open...
Scan...
OCR
Save Text
Clear Text Window
Remove Line Breaks
Copy all Texts to Clipboard
Export Text into Microsoft Word
Save Text
Quick OCR
Document
(i) 10-Year Defence Savings Certificates
Height: 1437 pixels Width: 971 pixels

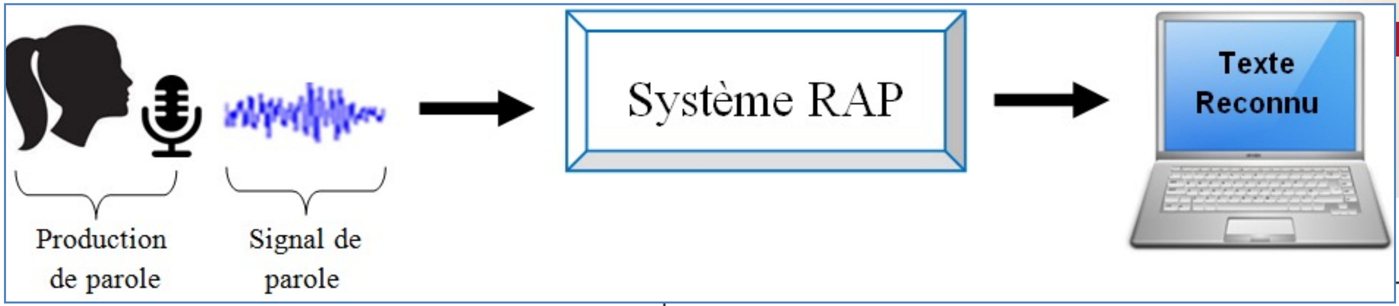
MODERN war means mechanised armies, tanks, aero-planes-thousands of them-and guns, in addition to men and munitions, and to equip and maintain a modern army for defending our homes will cost very large sums of money. Every patriodc Indian can help to make India strong and thus protect Hs home and family by saving as much as possible and lending his savings to Government. The Government of India have issued Defence Loans to suit the convenience of different classes of people. The object of these Defence Loans is to raisei money to expand and equip our army, navy and air force for the defence of our country. By subscribing to these loans you will be paying the best insurance for your freedom and happiness.

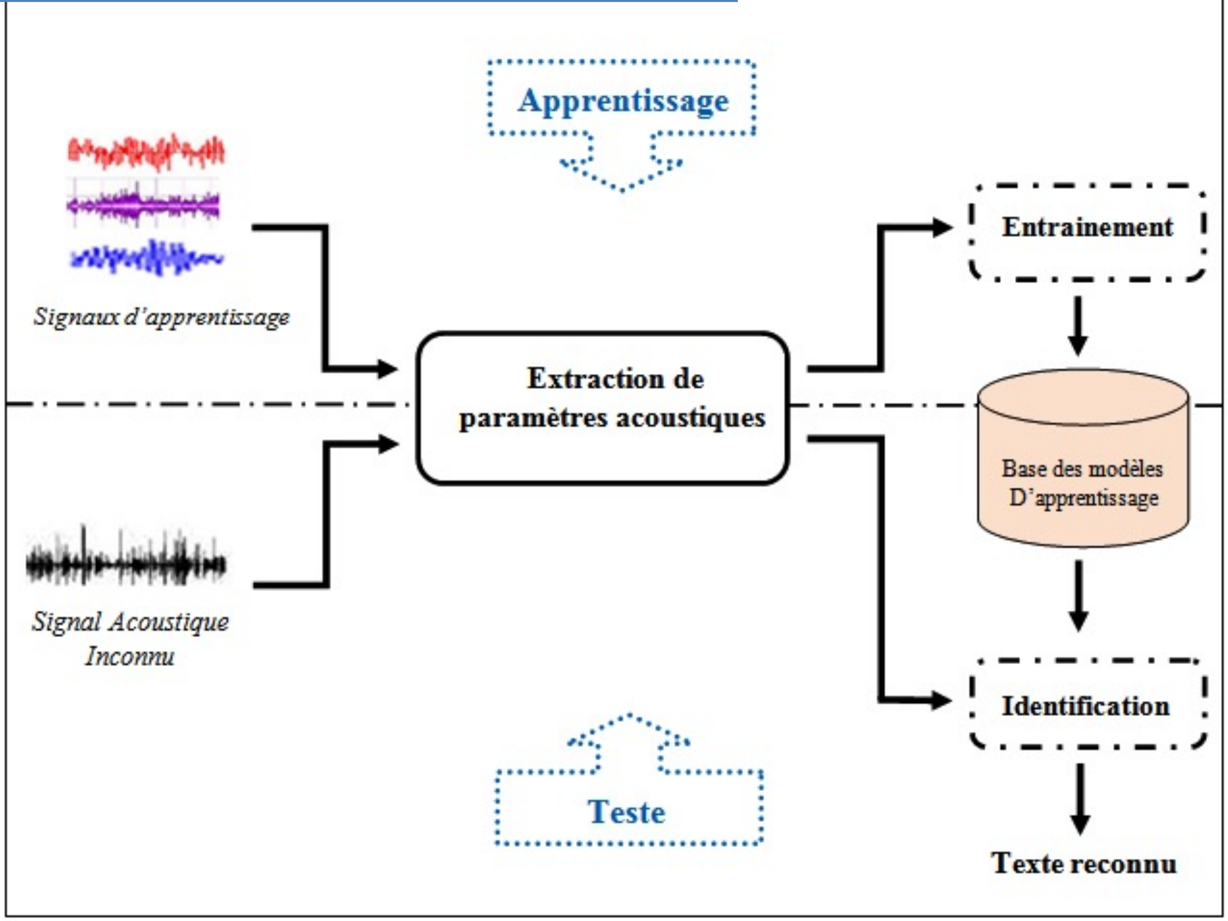
- **Sound** - Speech processing
  - speech recognition
  - speech synthesis



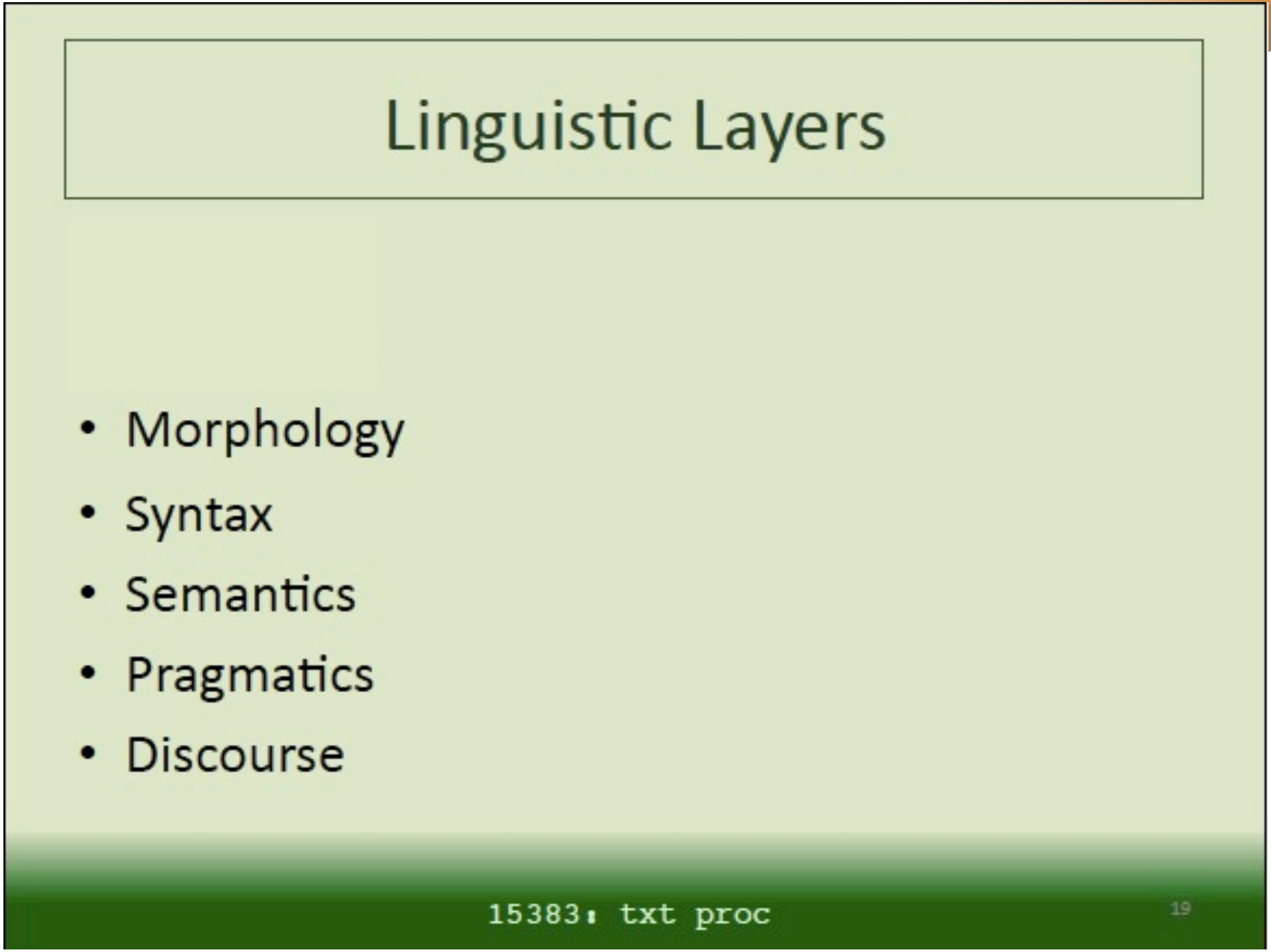
- **Text** - Text processing

# speech – using cmuSphinx




Apprentissage
Signaux d'apprentissage
Extraction de paramètres acoustiques
Entrainement
Base des modèles D'apprentissage
Signal Acoustique Inconnu
Identification
Teste
Texte reconnu

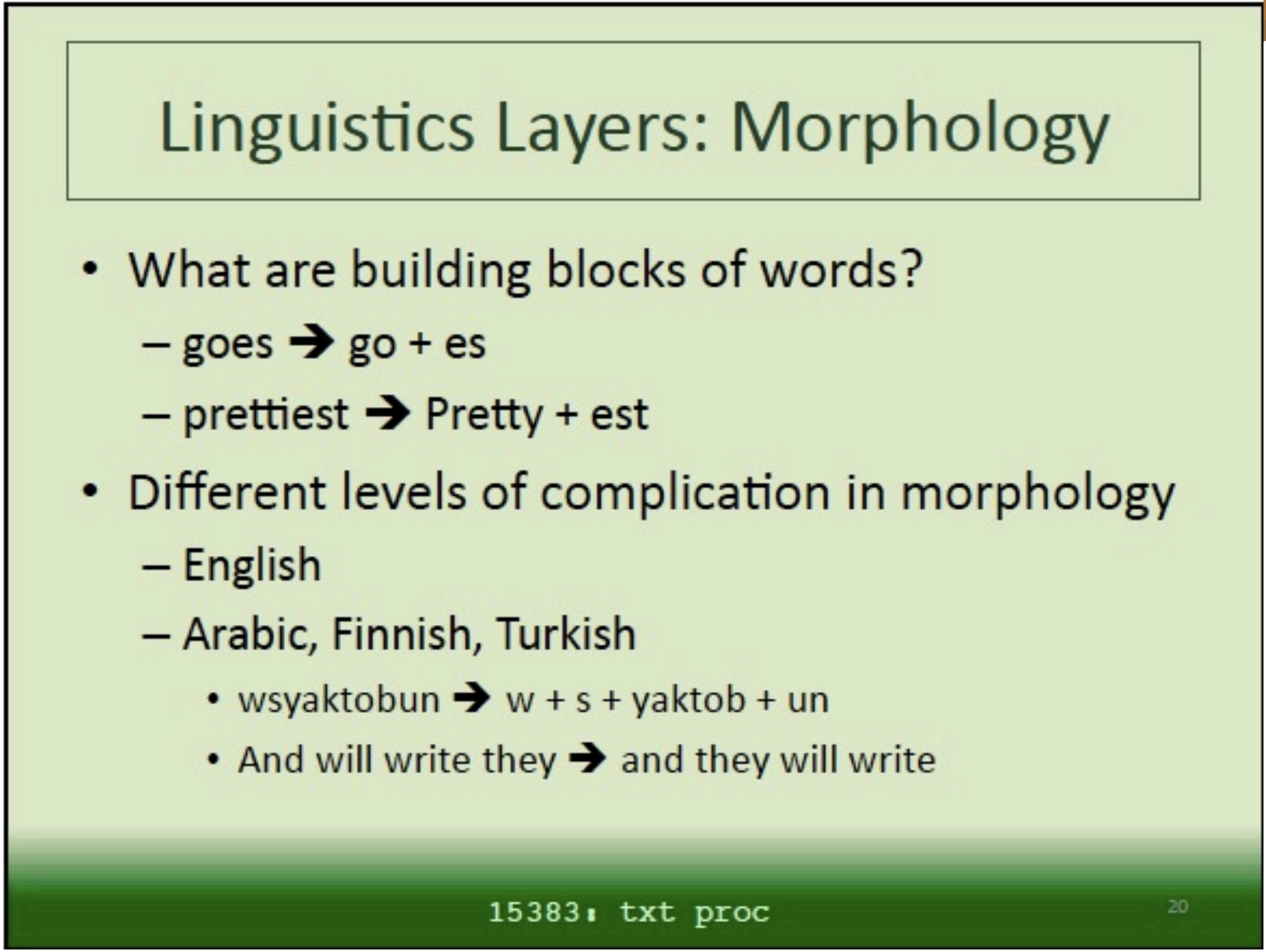
# *Levels for text*



## Linguistic Layers

- Morphology
- Syntax
- Semantics
- Pragmatics
- Discourse

# Levels for text

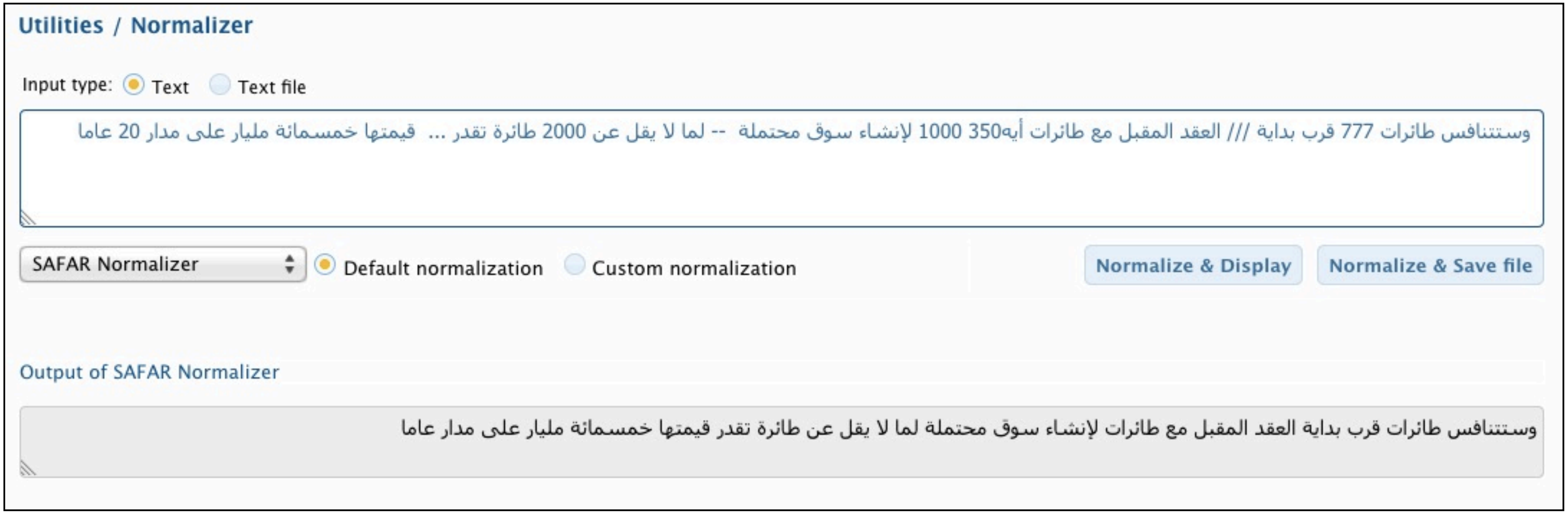


## Linguistics Layers: Morphology

- What are building blocks of words?
  - goes ➔ go + es
  - prettiest ➔ Pretty + est
- Different levels of complication in morphology
  - English
  - Arabic, Finnish, Turkish
    - wsyaktobun ➔ w + s + yaktob + un
    - And will write they ➔ and they will write

# *Basic text processing*

## Before Morphology - Normalizing



Input type: ● Text ○ Text file

وستتنافس طائرات 777 قرب بداية /// العقد المقبل مع طائرات أيه350 1000 لإنشاء سوق محتملة -- لما لا يقل عن 2000 طائرة تقدر ... قيمتها خمسمائة مليار على مدار 20 عاما

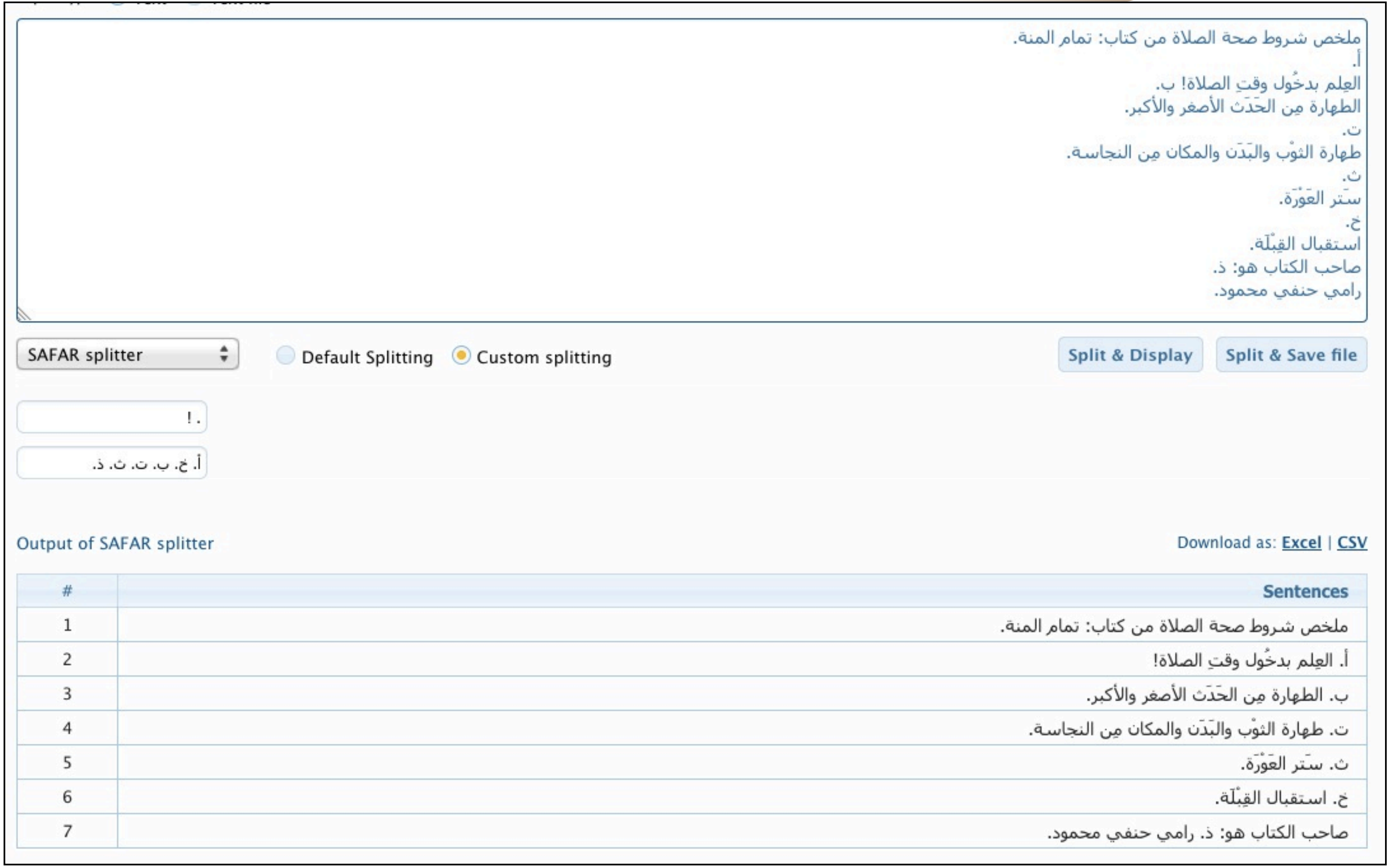
SAFAR Normalizer ● Default normalization ○ Custom normalization

Output of SAFAR Normalizer

وستتنافس طائرات قرب بداية العقد المقبل مع طائرات لإنشاء سوق محتملة لما لا يقل عن طائرة تقدر قيمتها خمسمائة مليار على مدار عاما

# Basic text processing

## Before Morphology - Splitting



ملخص شروط صحة الصلاة من كتاب: تمام المنة.
أ.
العِلم بدخُول وقتِ الصلاةَ! ب.
الطهارة مِن الحَدَث الأصغر والأكبر.
ت.
طهارة الثوْب والبَدَن والمكان مِن النجاسة.
ث.
سَتر العَوْرَة.
خ.
استقبال القِبْلَة.
صاحب الكتاب هو: ذ.
رامي حنفي محمود.

SAFAR splitter ○ Default Splitting ● Custom splitting

. !

أ. خ. ب. ت. ث. ذ.

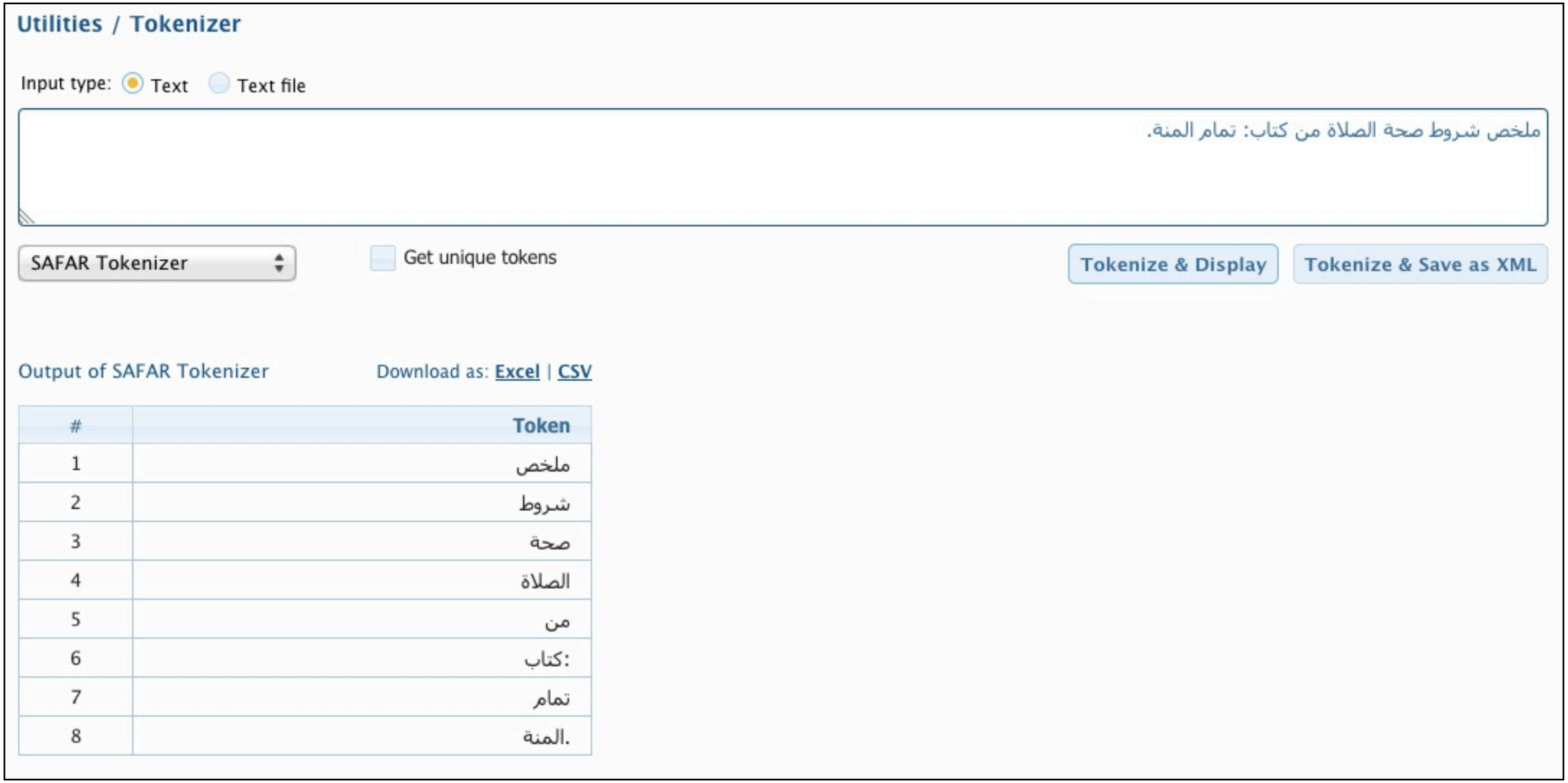
Output of SAFAR splitter

Download as: Excel | CSV

| # | Sentences |
|---|---|
| 1 | ملخص شروط صحة الصلاة من كتاب: تمام المنة. |
| 2 | أ. العِلم بدخُول وقتِ الصلاةَ! |
| 3 | ب. الطهارة مِن الحَدَث الأصغر والأكبر. |
| 4 | ت. طهارة الثوْب والبَدَن والمكان مِن النجاسة. |
| 5 | ث. سَتر العَوْرَة. |
| 6 | خ. استقبال القِبْلَة. |
| 7 | صاحب الكتاب هو: ذ. رامي حنفي محمود. |

# *Basic text processing*

## Before Morphology – Tokenizing



Utilities / Tokenizer

Input type: ● Text ○ Text file

ملخص شروط صحة الصلاة من كتاب: تمام المنة.

SAFAR Tokenizer

□ Get unique tokens

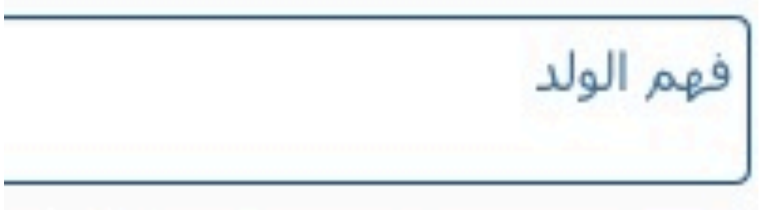
Output of SAFAR Tokenizer

Download as: Excel | CSV

| # | Token |
|---|---|
| 1 | ملخص |
| 2 | شروط |
| 3 | صحة |
| 4 | الصلاة |
| 5 | من |
| 6 | كتاب: |
| 7 | تمام |
| 8 | المنة. |

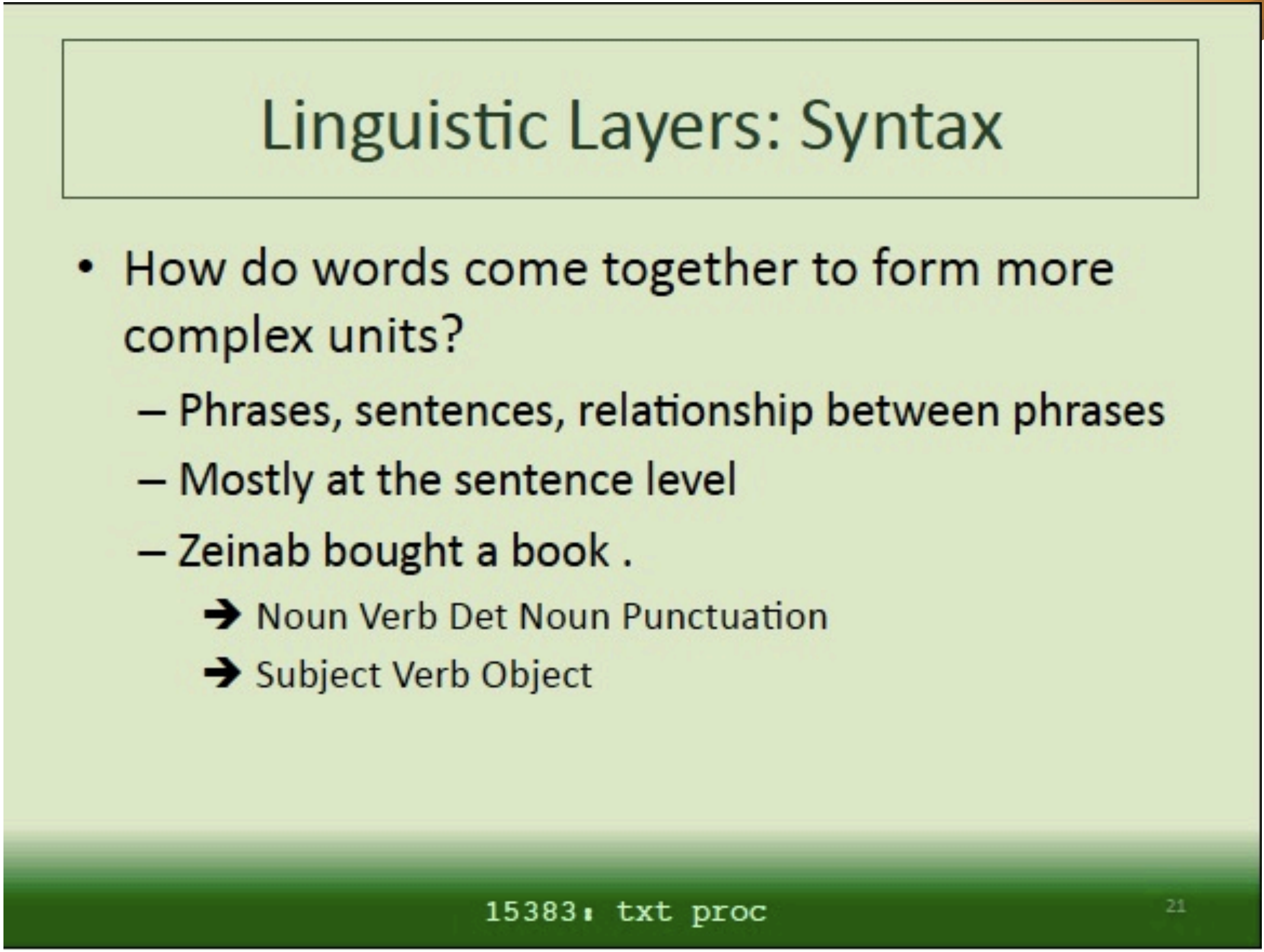
# Morphology

- Morphological analysis (lexical process): it is the study of the structure of words. It specifies how words are constructed by identifying lexical components and their properties
- Ambiguity
  - Ex: it lights (noun, verb, adjective)



فهم الولد

# *Levels for text*

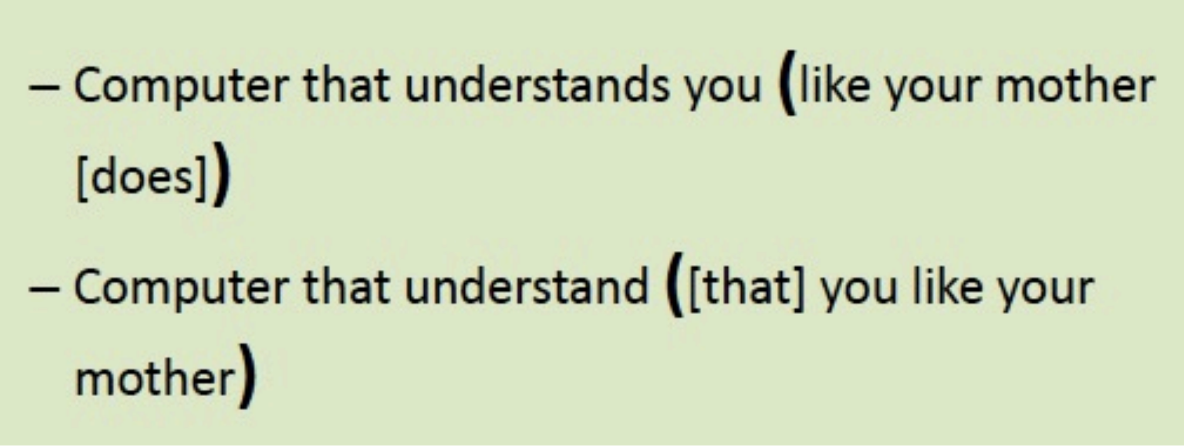


## Linguistic Layers: Syntax

- How do words come together to form more complex units?
  - Phrases, sentences, relationship between phrases
  - Mostly at the sentence level
  - Zeinab bought a book .
    - ➔ Noun Verb Det Noun Punctuation
    - ➔ Subject Verb Object

# Syntax

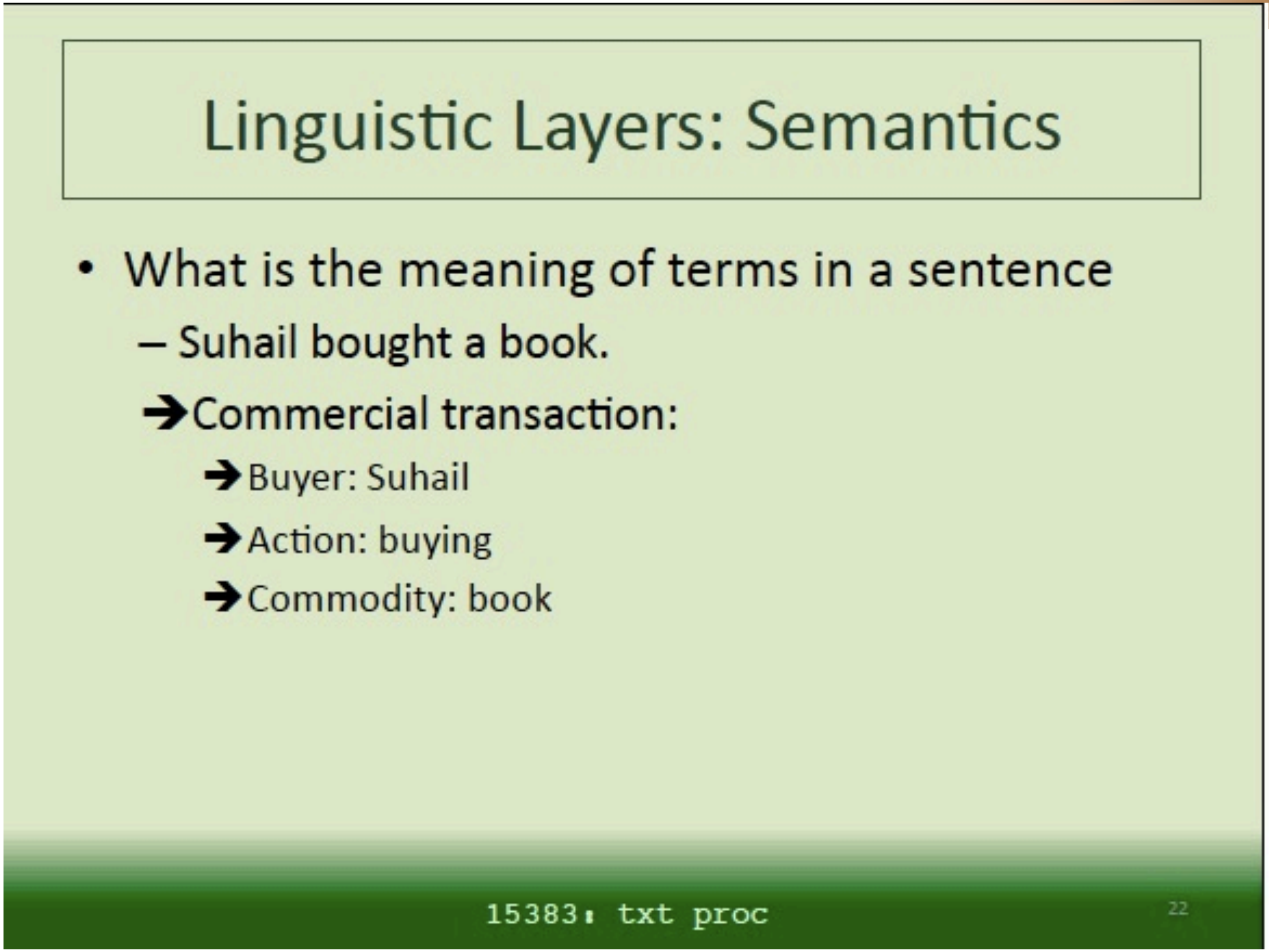
- Syntactic Analysis: Treats the way words can combine to form sentences. It allows to identify the structure of the sentence and the links between the words
- Ambiguity:



– Computer that understands you (like your mother [does])

– Computer that understand ([that] you like your mother)

# Levels for text

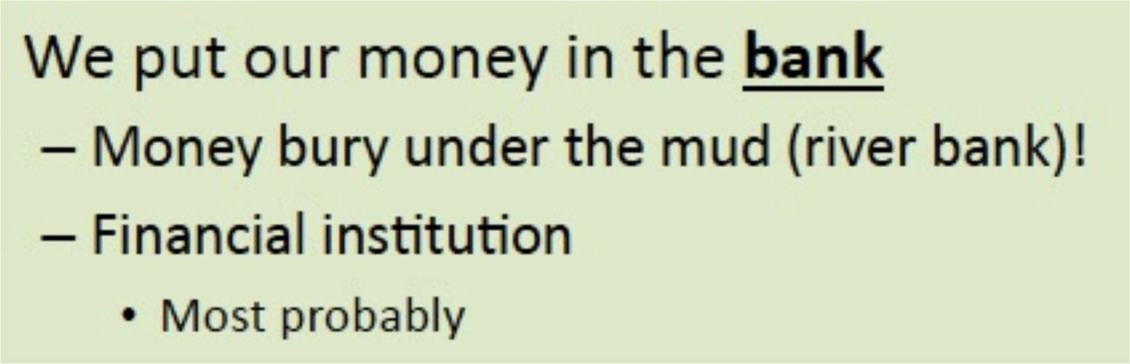


## Linguistic Layers: Semantics

- What is the meaning of terms in a sentence
  - – Suhail bought a book.
  - ➔Commercial transaction:
    - ➔Buyer: Suhail
    - ➔Action: buying
    - ➔Commodity: book

# Semantics

- Semantic analysis: it identifies the meaning of the phrase outside the context (to be able to translate it for instance)
- Ambiguity:



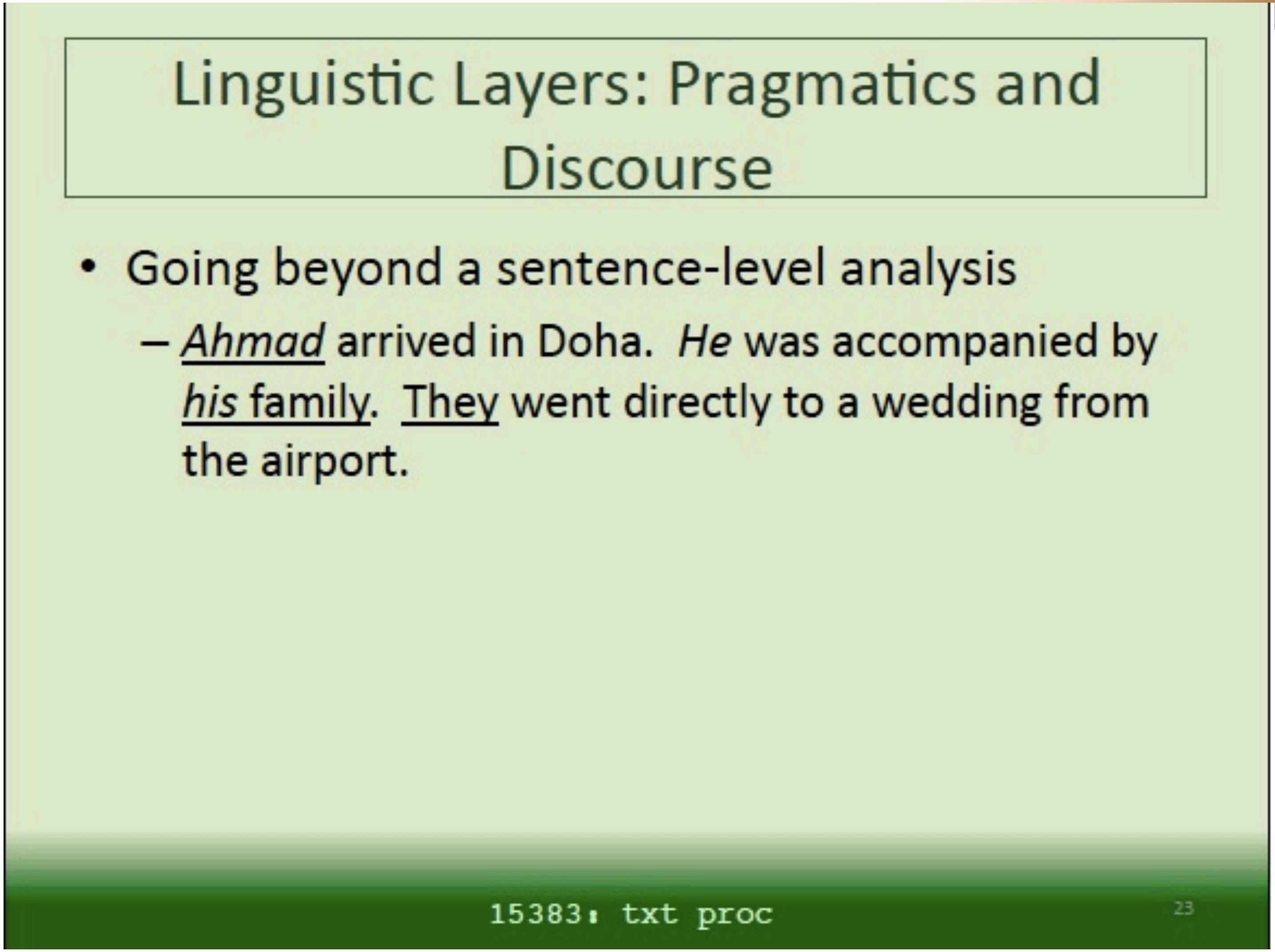
We put our money in the **bank**

– Money bury under the mud (river bank)!

– Financial institution

  - Most probably

# Levels for text



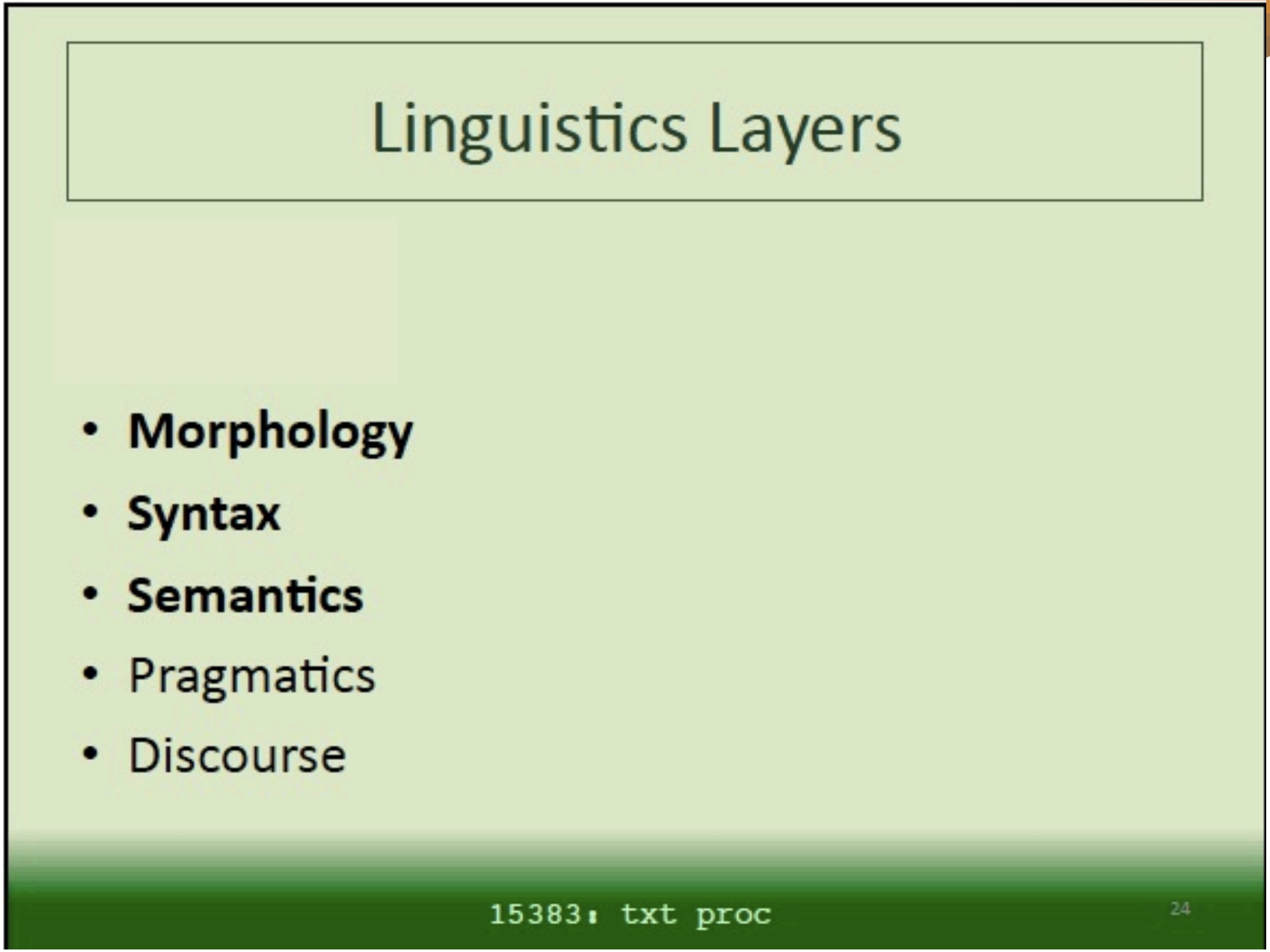
## Linguistic Layers: Pragmatics and Discourse

- Going beyond a sentence-level analysis
  - *Ahmad* arrived in Doha. *He* was accompanied by *his* family. They went directly to a wedding from the airport.

# Pragmatics

- Pragmatic analysis: it aims to study the meaning of the sentence in the context. It makes it possible to find the real meaning of sentences related to situational and contextual conditions

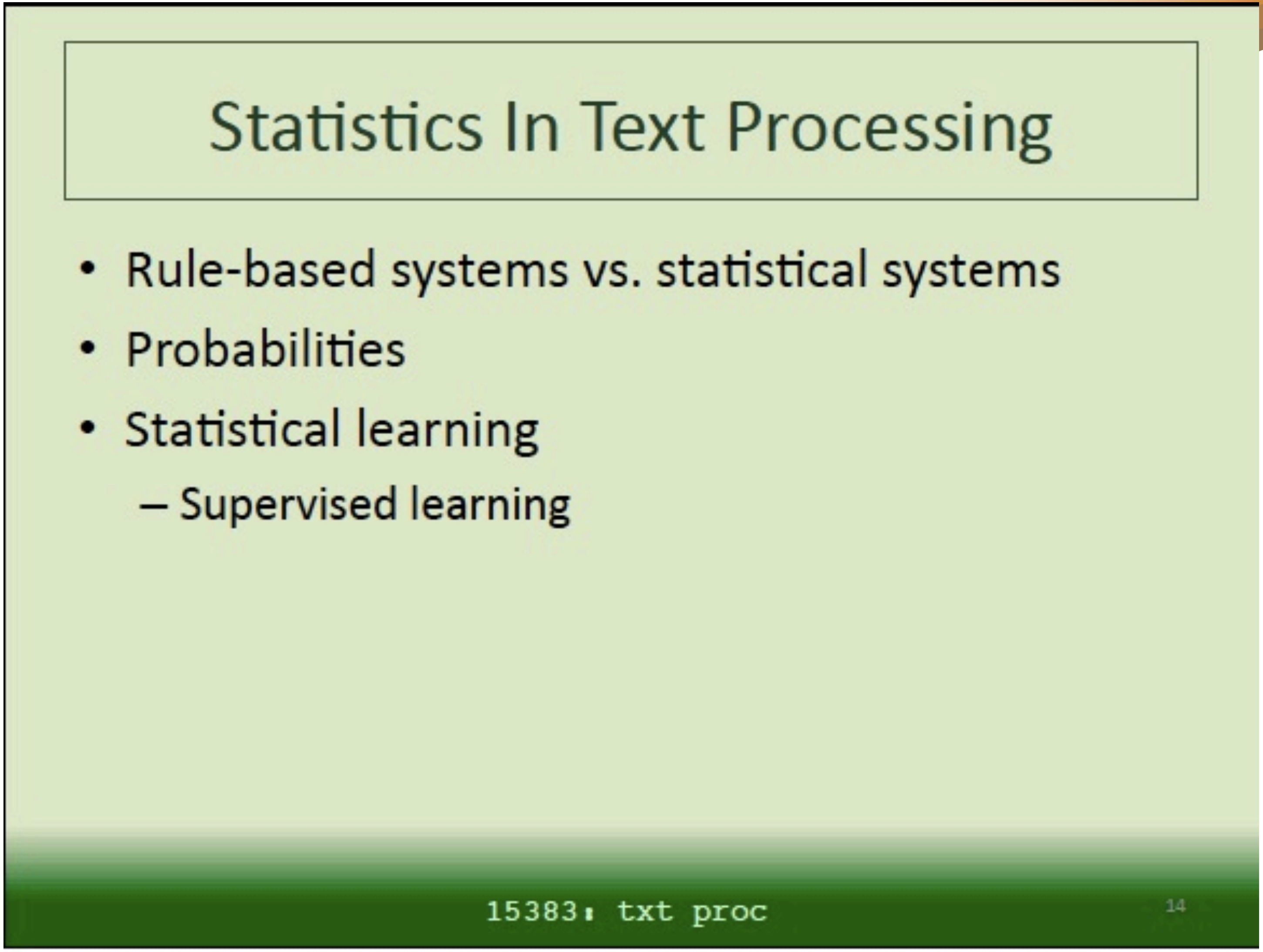
# *Levels for text*



## Linguistics Layers

- **Morphology**
- **Syntax**
- **Semantics**
- Pragmatics
- Discourse

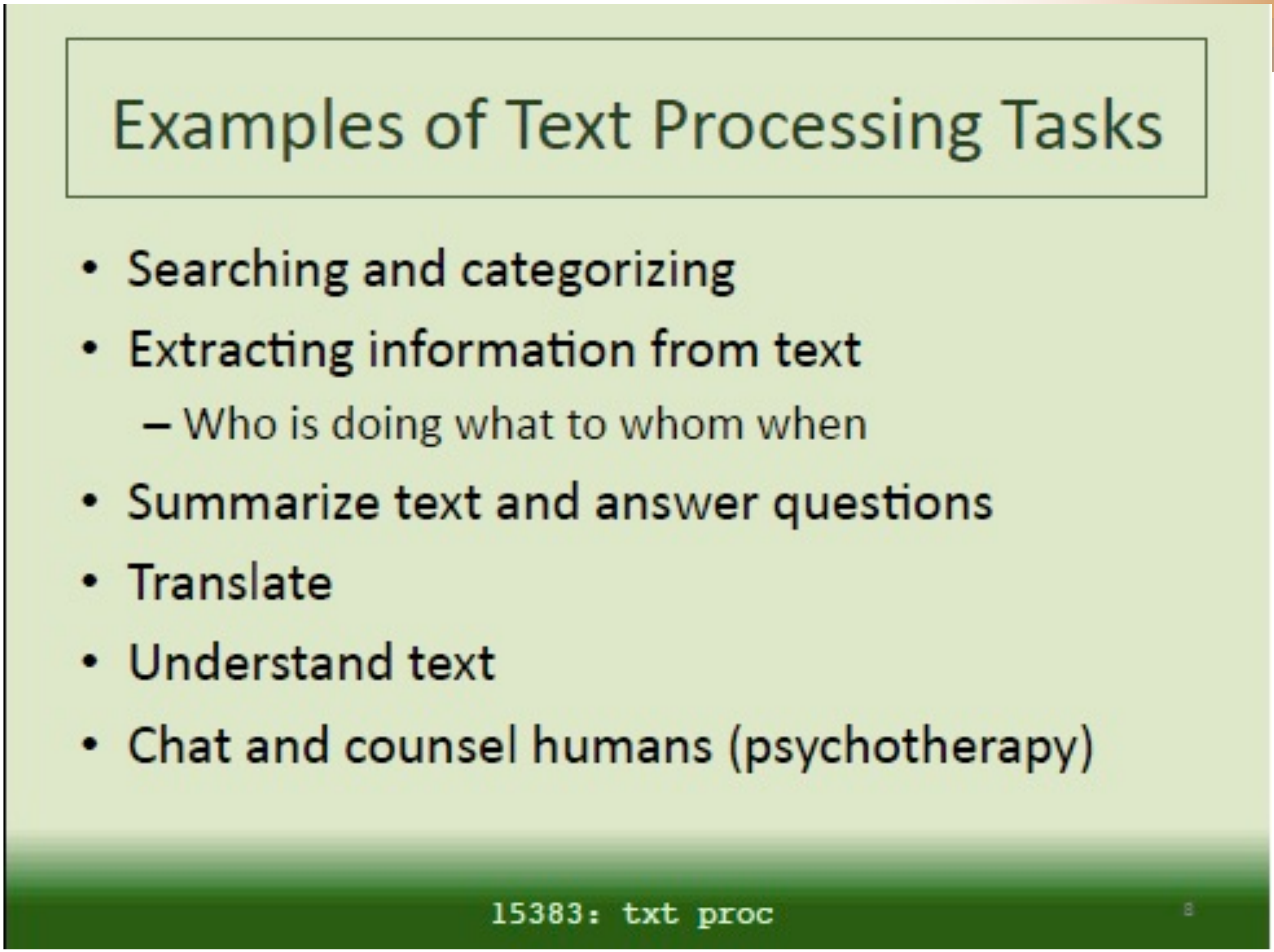
# Applications – Rules or Stats



## Statistics In Text Processing

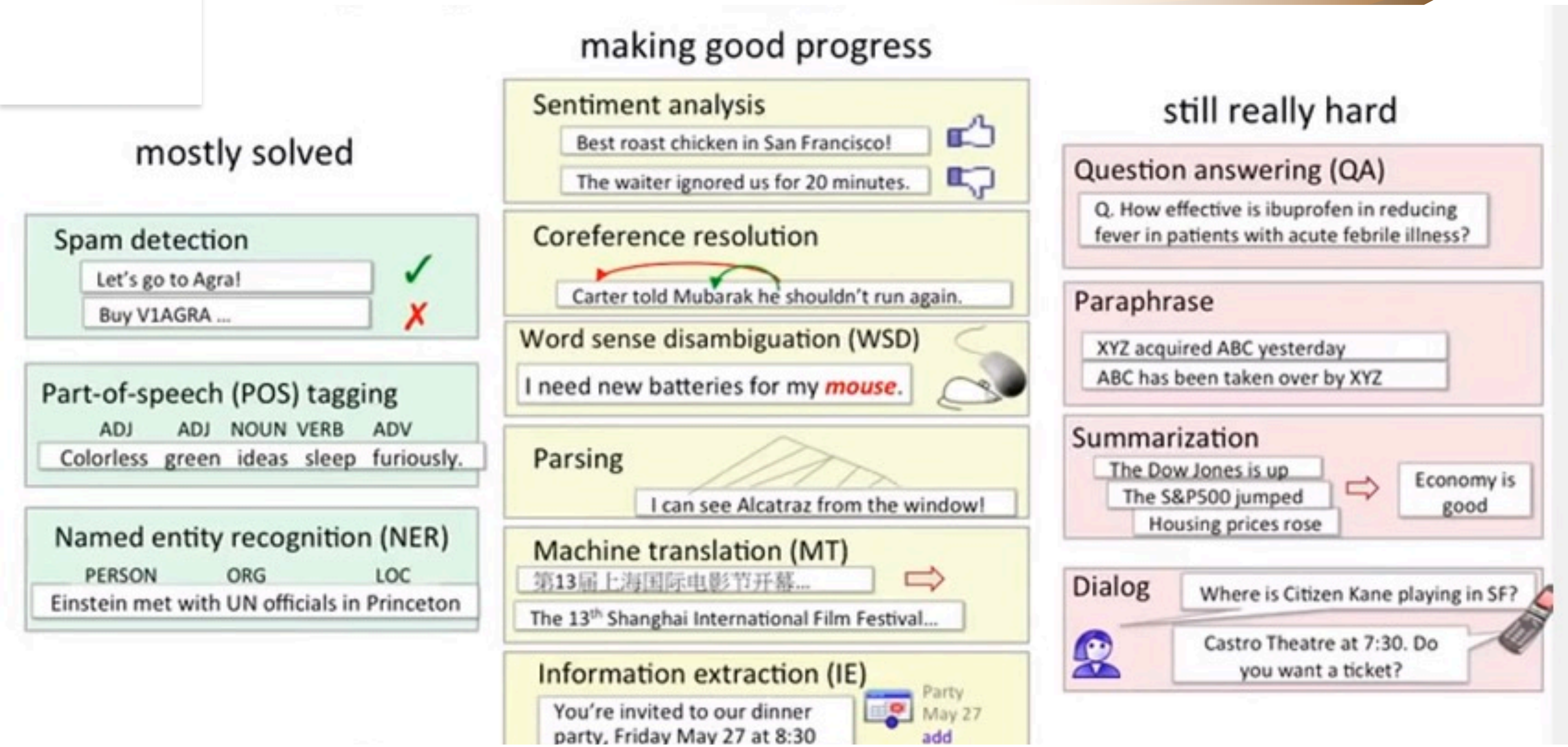
- Rule-based systems vs. statistical systems
- Probabilities
- Statistical learning
  - – Supervised learning

# *Applications*



## Examples of Text Processing Tasks

- Searching and categorizing
- Extracting information from text
  - Who is doing what to whom when
- Summarize text and answer questions
- Translate
- Understand text
- Chat and counsel humans (psychotherapy)



## mostly solved

### Spam detection

Let's go to Agra! ✓

Buy V1AGRA … ✗

### Part-of-speech (POS) tagging

| ADJ | ADJ | NOUN | VERB | ADV |
|---|---|---|---|---|
| Colorless | green | ideas | sleep | furiously. |

### Named entity recognition (NER)

| PERSON | | | ORG | | LOC |
|---|---|---|---|---|---|
| Einstein | met | with | UN officials | in | Princeton |

## making good progress

### Sentiment analysis

Best roast chicken in San Francisco! 👍

The waiter ignored us for 20 minutes. 👎

### Coreference resolution

Carter told Mubarak he shouldn't run again.

### Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

### Parsing

I can see Alcatraz from the window!

### Machine translation (MT)

第13届上海国际电影节开幕…

⇨

The 13th Shanghai International Film Festival…

### Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

## still really hard

### Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

### Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

### Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

⇨

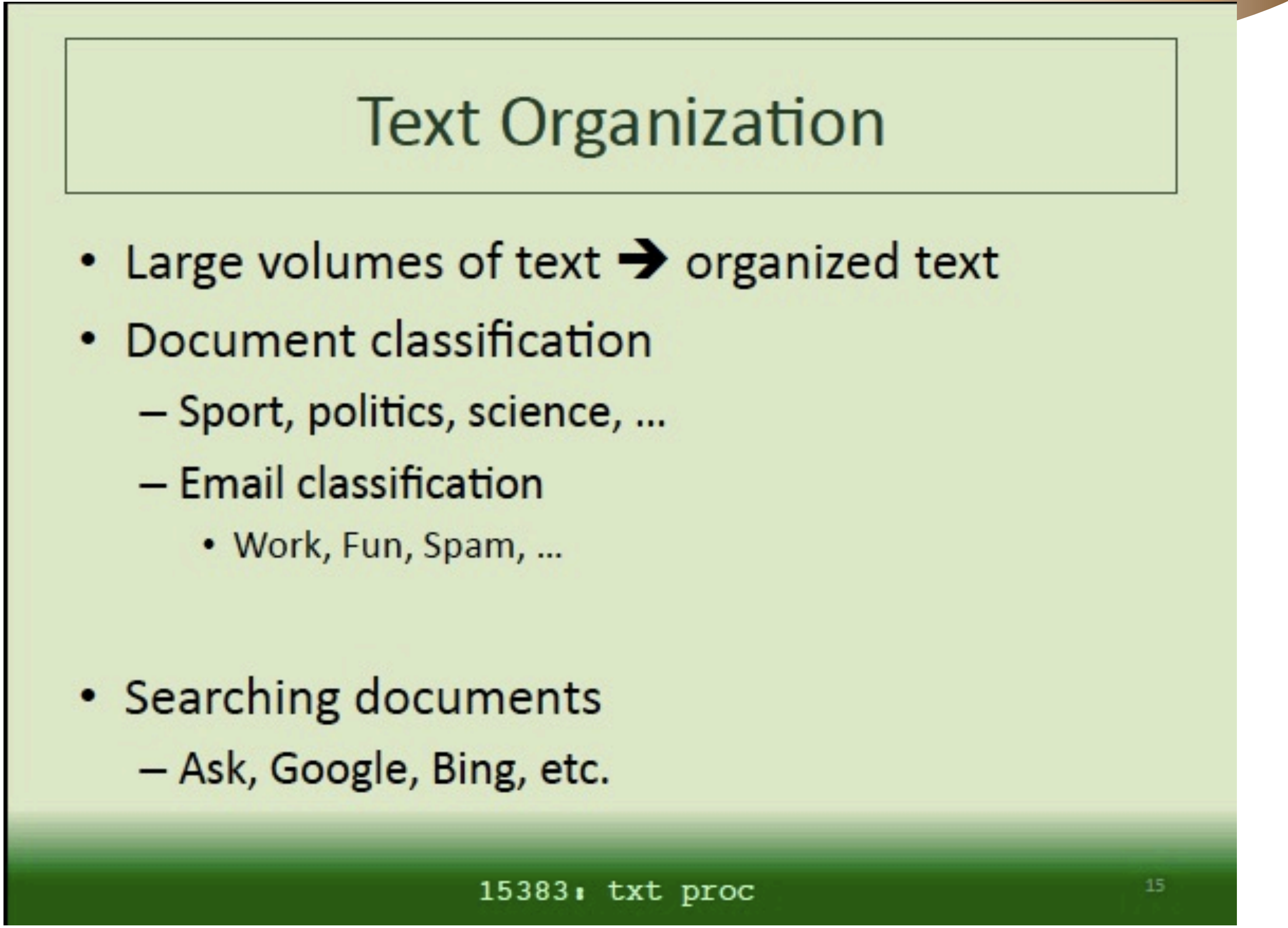
Economy is good

### Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

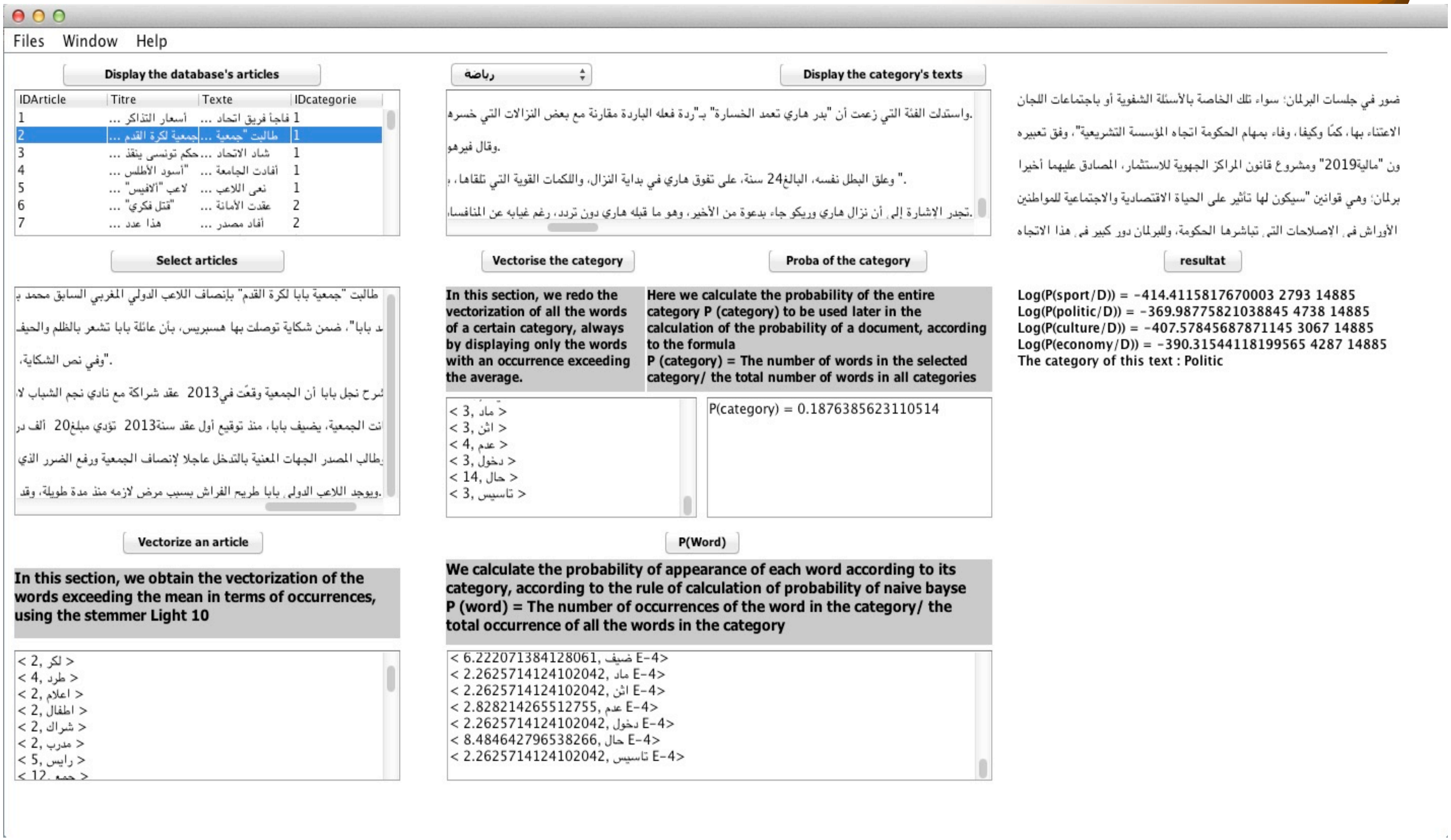
# Applications



## Text Organization

- Large volumes of text ➔ organized text
- Document classification
  - Sport, politics, science, …
  - Email classification
    - Work, Fun, Spam, …
- Searching documents
  - Ask, Google, Bing, etc.

# appl / services – classification



| IDArticle | Titre | Texte | IDcategorie |
|---|---|---|---|
| 1 | أسعار التذاكر ... | فاجأ فريق اتحاد ... | 1 |
| 2 | جمعية لكرة القدم ... | طالبت "جمعية ... | 1 |
| 3 | حكم تونسي ينقذ ... | شاد الاتحاد ... | 1 |
| 4 | "أسود الأطلس ... | أفادت الجامعة ... | 1 |
| 5 | لاعب "ألافيس" ... | نعى اللاعب ... | 1 |
| 6 | "قتل فكري" ... | عقدت الأمانة ... | 2 |
| 7 | هذا عدد ... | أفاد مصدر ... | 2 |

واستدلت الفئة التي زعمت أن "بدر هاري تعمد الخسارة" بـ"ردة فعله الباردة مقارنة مع بعض النزالات التي خسره.

وقال فيرهو.

" وعلق البطل نفسه، البالغ24 سنة، على تفوق هاري في بداية النزال، واللكمات القوية التي تلقاها، ب.

تجدر الإشارة إلى أن نزال هاري وريكو جاء بدعوة من الأخير، وهو ما قبله هاري دون تردد، رغم غيابه عن المنافسا.

ضور في جلسات البرلمان؛ سواء تلك الخاصة بالأسئلة الشفوية أو باجتماعات اللجان

الاعتناء بها، كمًا وكيفا، وفاء بمهام الحكومة اتجاه المؤسسة التشريعية"، وفق تعبيره

ون "مالية2019" ومشروع قانون المراكز الجهوية للاستثمار، المصادق عليهما أخيرا

برلمان؛ وهي قوانين "سيكون لها تأثير على الحياة الاقتصادية والاجتماعية للمواطنين

الأوراش في الإصلاحات التي تباشرها الحكومة، وللبرلمان دور كبير في هذا الاتجاه

طالبت "جمعية بابا لكرة القدم" بإنصاف اللاعب الدولي المغربي السابق محمد ب

د بابا"، ضمن شكاية توصلت بها هسبريس، بأن عائلة بابا تشعر بالظلم والحيف

وفي نص الشكاية،".

شرح نجل بابا أن الجمعية وقعّت في2013 عقد شراكة مع نادي نجم الشباب لا

انت الجمعية، يضيف بابا، منذ توقيع أول عقد سنة2013 تؤدي مبلغ20 ألف در

يطالب المصدر الجهات المعنية بالتدخل عاجلا لإنصاف الجمعية ورفع الضرر الذي

ويوجد اللاعب الدولي بابا طريح الفراش بسبب مرض لازمه منذ مدة طويلة، وقد.

In this section, we redo the vectorization of all the words of a certain category, always by displaying only the words with an occurrence exceeding the average.

Here we calculate the probability of the entire category P (category) to be used later in the calculation of the probability of a document, according to the formula
P (category) = The number of words in the selected category/ the total number of words in all categories

```
< 3, ماد >
< 3, اثن >
< 4, عدم >
< 3, دخول >
< 14, حال >
< 3, تاسيس >
```

```
P(category) = 0.1876385623110514
```

```
Log(P(sport/D)) = -414.4115817670003 2793 14885
Log(P(politic/D)) = -369.98775821038845 4738 14885
Log(P(culture/D)) = -407.57845687871145 3067 14885
Log(P(economy/D)) = -390.31544118199565 4287 14885
The category of this text : Politic
```

In this section, we obtain the vectorization of the words exceeding the mean in terms of occurrences, using the stemmer Light 10

```
< 2, لكر >
< 4, طرد >
< 2, اعلام >
< 2, اطفال >
< 2, شراك >
< 2, مدرب >
< 5, رايس >
< 12, جمع >
```

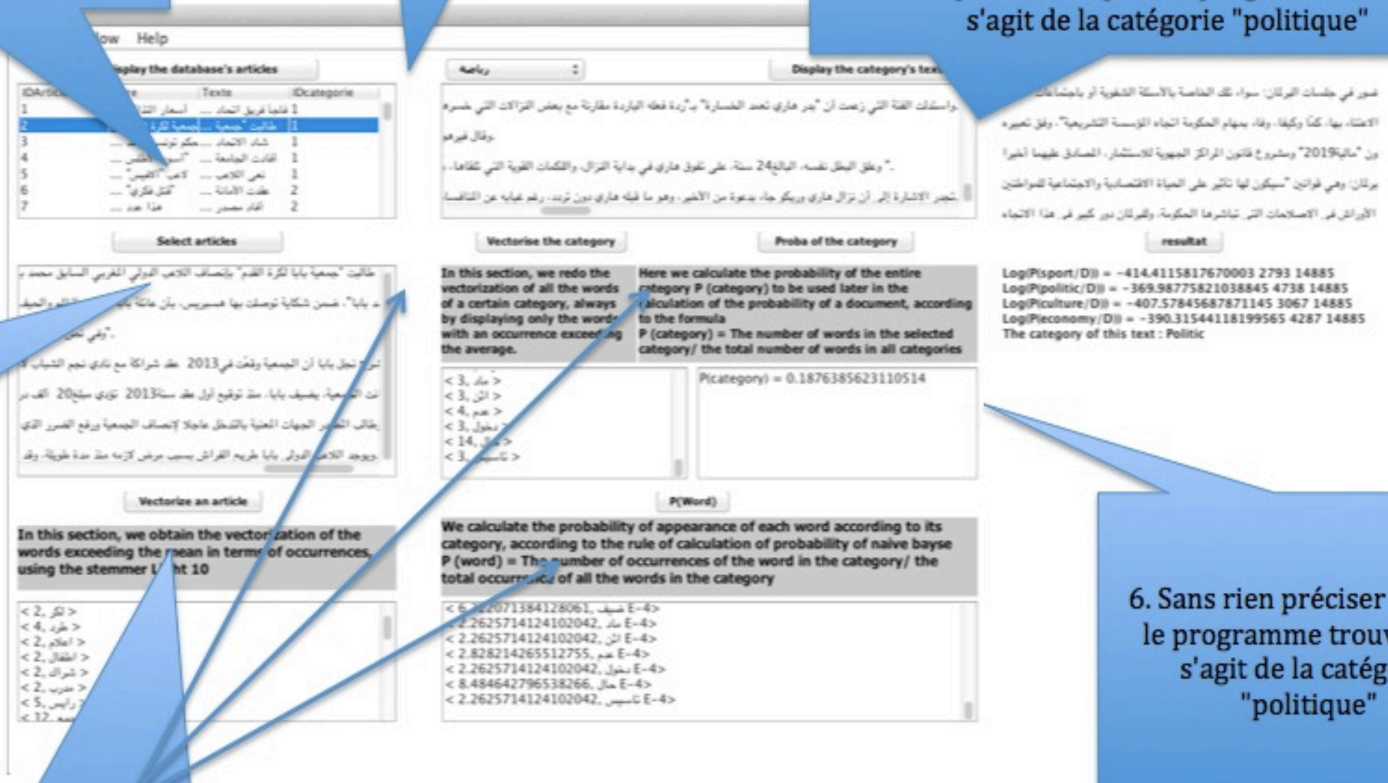
We calculate the probability of appearance of each word according to its category, according to the rule of calculation of probability of naive bayse
P (word) = The number of occurrences of the word in the category/ the total occurrence of all the words in the category

```
< 6.222071384128061, ضيف E-4>
< 2.2625714124102042, ماد E-4>
< 2.2625714124102042, اثن E-4>
< 2.828214265512755, عدم E-4>
< 2.2625714124102042, دخول E-4>
< 8.484642796538266, حال E-4>
< 2.2625714124102042, تاسيس E-4>
```

1. Au préablable, nous avons dumpé tous les articles hespress durant une période donnée en 2017. Ce tableau affiche tous ces articles avec leur numéro de catégorie. Les catégories considérées sont: sport, politique, culture et économie

4. Ici par exemple catégorie sport

5. voici ensuite la partie qu'un end user peut exploiter en mettant son texte et il demande au programme de la catégoriser automatiquement. j'ai pris par exemple un article de hespress daté du 13 déc 2018 https://www.hespress.com/politique/415347.html. sans rien préciser de plus, le programme trouve qu'il s'agit de la catégorie "politique"



| IDArtic | re | Texte | IDcategorie |
|---|---|---|---|
| 1 | أسعار الثا... | فاجأ فريق اتحاد ... | 1 |
| 2 | جمعية لكرة ... | طالبت "جمعية ... | 1 |
| 3 | ... حكم تونسي | شاد الاتحاد ... | 1 |
| 4 | ... "أسود الأطلس" | أقالت الجامعة ... | 1 |
| 5 | ... لاعب "الكليس" | نعى اللاعب ... | 1 |
| 6 | ... "قتل فكري" | عقدت الأمانة ... | 2 |
| 7 | ... هذا عدد | أفاد مصدر ... | 2 |

طالبت "جمعية بابا لكرة القدم" بإنصاف اللاعب الدولي المغربي السابق محمد

بابا"، ضمن شكاية توصلت بها هسبريس، بأن عائلة

"وفي نفس

2013 عقد شراكة مع نادي نجم الشباب

الجمعية، يضيف بابا، منذ توقيع أول عقد سنة 2013 تؤدي مبلغ 20 ألف در

وطالب القاصر الجهات المعنية بالتدخل عاجلا لإنصاف الجمعية ورفع الضرر الذي

ويوجد اللاعب الدولي بابا طريح الفراش بسبب مرض لازمه منذ مدة طويلة، وقد

In this section, we obtain the vectorization of the words exceeding the mean in terms of occurrences, using the stemmer Light 10

```
< 2, لكر >
< 4, طرد >
< 2, اعلام >
< 2, اطفال >
< 2, شراك >
< 2, مدرب >
< 5, رايس >
< 12,
```

واستنكرت الفئة التي زعمت أن "بدر هاري تعمد الخسارة" بـ"ردة فعله الباردة مقارنة مع بعض النزالات التي خسرها

وقال فيرهو

"وعلق البطل نفسه، البالغ 24 سنة، على تفوق هاري في بداية النزال، واللكمات القوية التي تلقاها.

تجدر الإشارة إلى أن نزال هاري وريكو جاء بدعوة من الأخير، وهو ما قبله هاري دون تردد، رغم غيابه عن المنافسة

In this section, we redo the vectorization of all the words of a certain category, always by displaying only the words with an occurrence exceeding the average.

```
< 3, ساد >
< 3, اثن >
< 4, عدم >
< 3, دخول >
< 14,
< 3, تأسيس >
```

Here we calculate the probability of the entire category P (category) to be used later in the calculation of the probability of a document, according to the formula
P (category) = The number of words in the selected category/ the total number of words in all categories

```
P(category) = 0.1876385623110514
```

We calculate the probability of appearance of each word according to its category, according to the rule of calculation of probability of naive bayse
P (word) = The number of occurrences of the word in the category/ the total occurrences of all the words in the category

```
< 6.222071384128061, ضيف E-4>
< 2.2625714124102042, ساد E-4>
< 2.2625714124102042, اثن E-4>
< 2.828214265512755, عدم E-4>
< 2.2625714124102042, دخول E-4>
< 8.484642796538266, حال E-4>
< 2.2625714124102042, تأسيس E-4>
```

حضور في جلسات البرلمان؛ سواء تلك الخاصة بالأسئلة الشفوية أو باجتماعات

الاعتناء بها، كماً وكيفاً، وفاء بمهام الحكومة اتجاه المؤسسة التشريعية"، وفق تعبيره

ون "مالية 2019" ومشروع قانون المراكز الجهوية للاستثمار، المصادق عليهما أخيرا

برلمان؛ وهي قوانين "سيكون لها تأثير على الحياة الاقتصادية والاجتماعية للمواطنين

الأوراش في الاصلاحات التي تباشرها الحكومة، وللبرلمان دور كبير في هذا الاتجاه

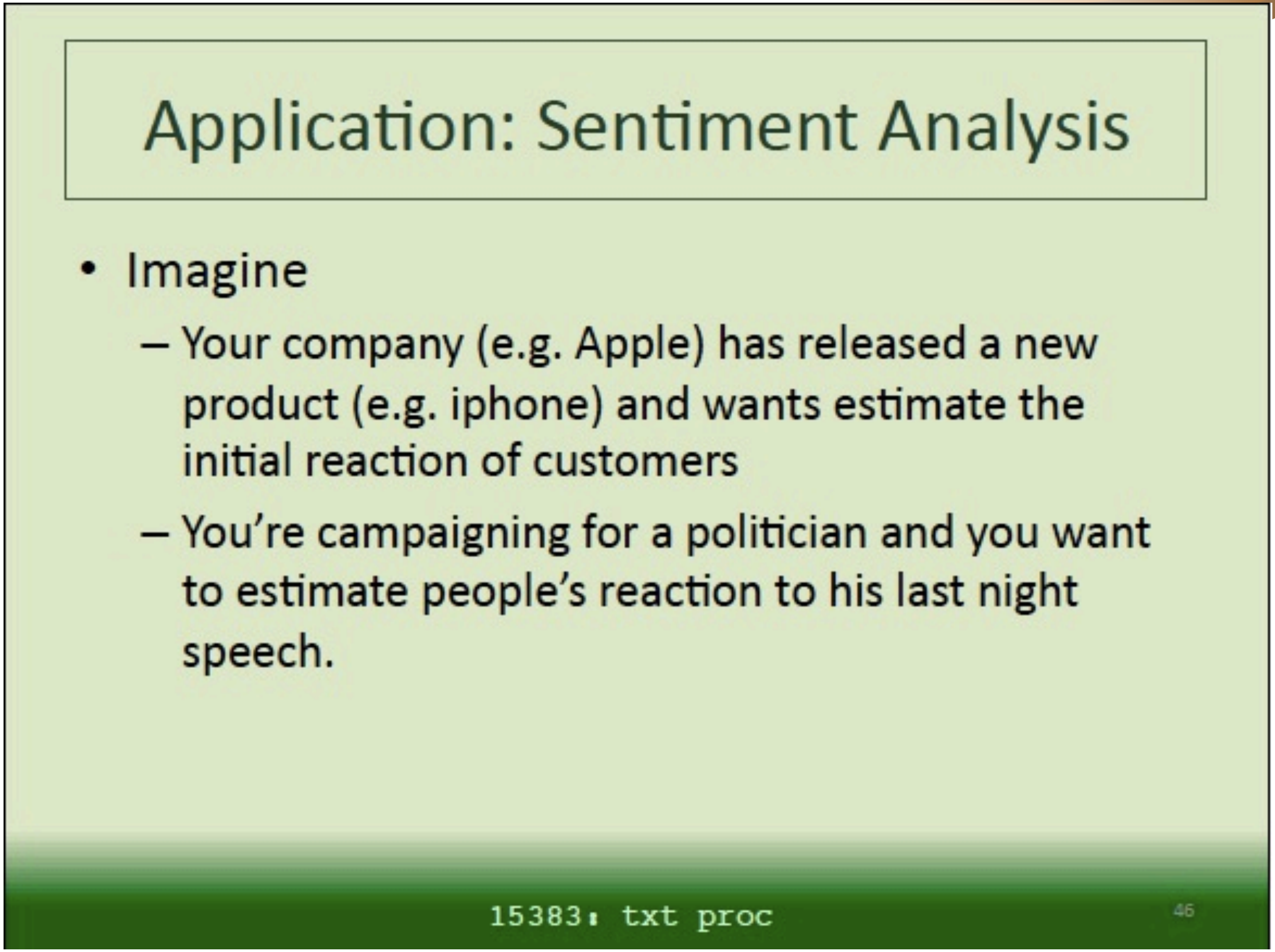
```
Log(P(sport/D)) = -414.4115817670003 2793 14885
Log(P(politic/D)) = -369.98775821038845 4738 14885
Log(P(culture/D)) = -407.57845687871145 3067 14885
Log(P(economy/D)) = -390.31544118199565 4287 14885
The category of this text : Politic
```

2. Il est possible ensuite de cliquer sur un article donné et afficher son contenu en cliquant sur "select articles"

6. Sans rien préciser de plus, le programme trouve qu'il s'agit de la catégorie "politique"

3. nous faisons ensuite un process de vectorisation et de calcul probabiliste pour que l'ordinateur apprenne et modélise toutes les catégories.

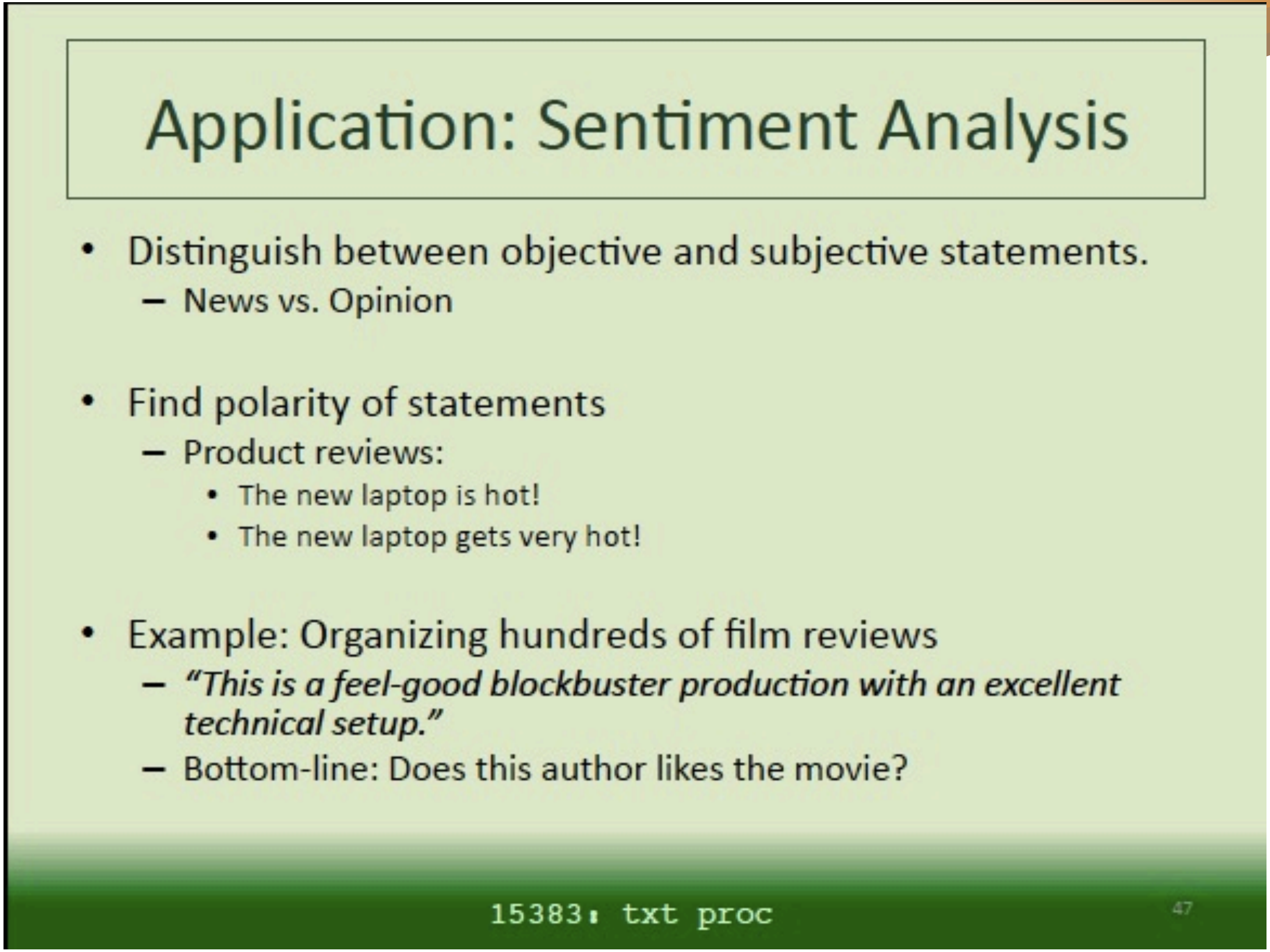
# *Applications*



## Application: Sentiment Analysis

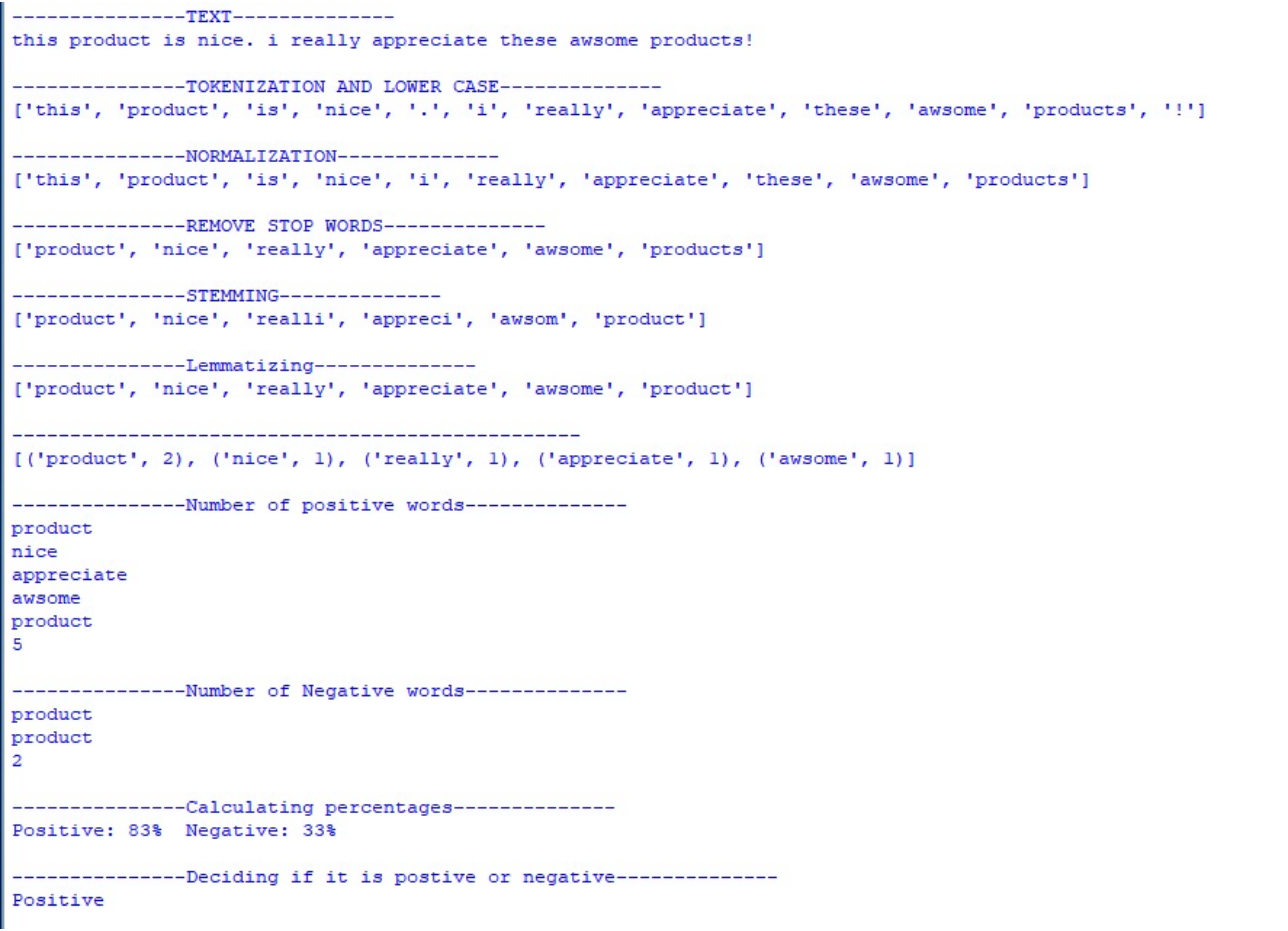
- Imagine
  - Your company (e.g. Apple) has released a new product (e.g. iphone) and wants estimate the initial reaction of customers
  - You're campaigning for a politician and you want to estimate people's reaction to his last night speech.

# Applications

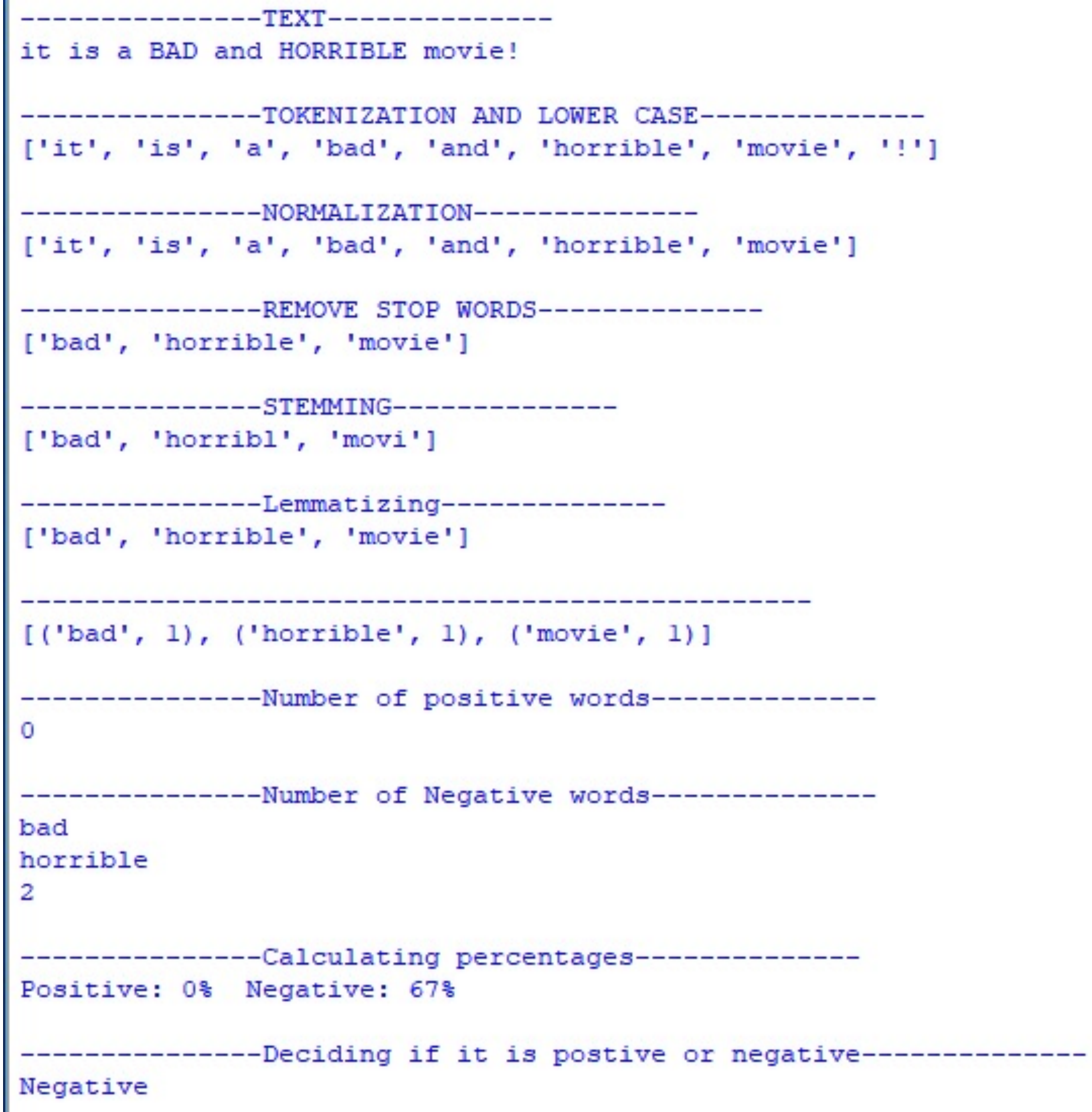


## Application: Sentiment Analysis

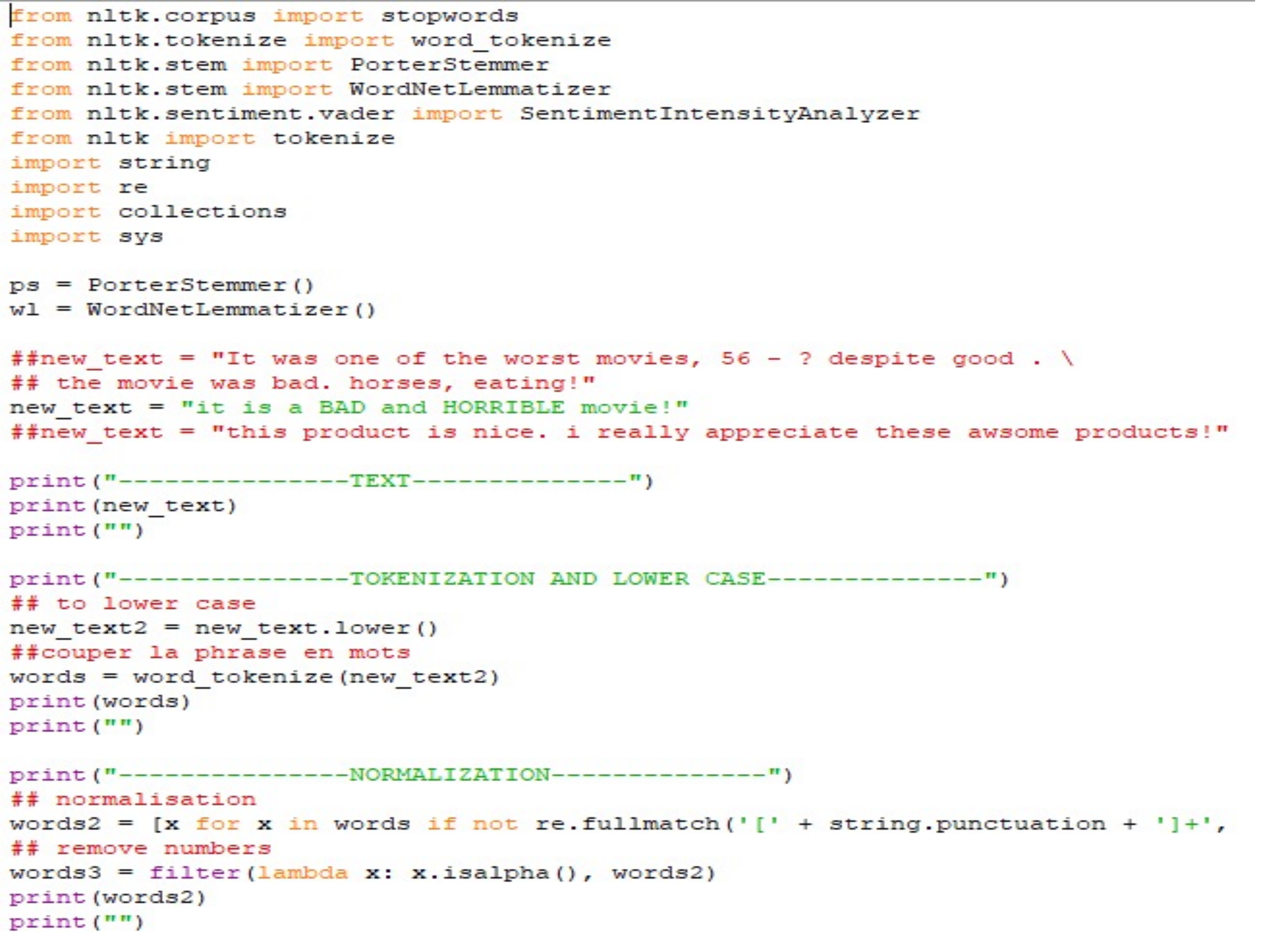
- Distinguish between objective and subjective statements.
  - News vs. Opinion
- Find polarity of statements
  - Product reviews:
    - The new laptop is hot!
    - The new laptop gets very hot!
- Example: Organizing hundreds of film reviews
  - *“This is a feel-good blockbuster production with an excellent technical setup.”*
  - Bottom-line: Does this author likes the movie?



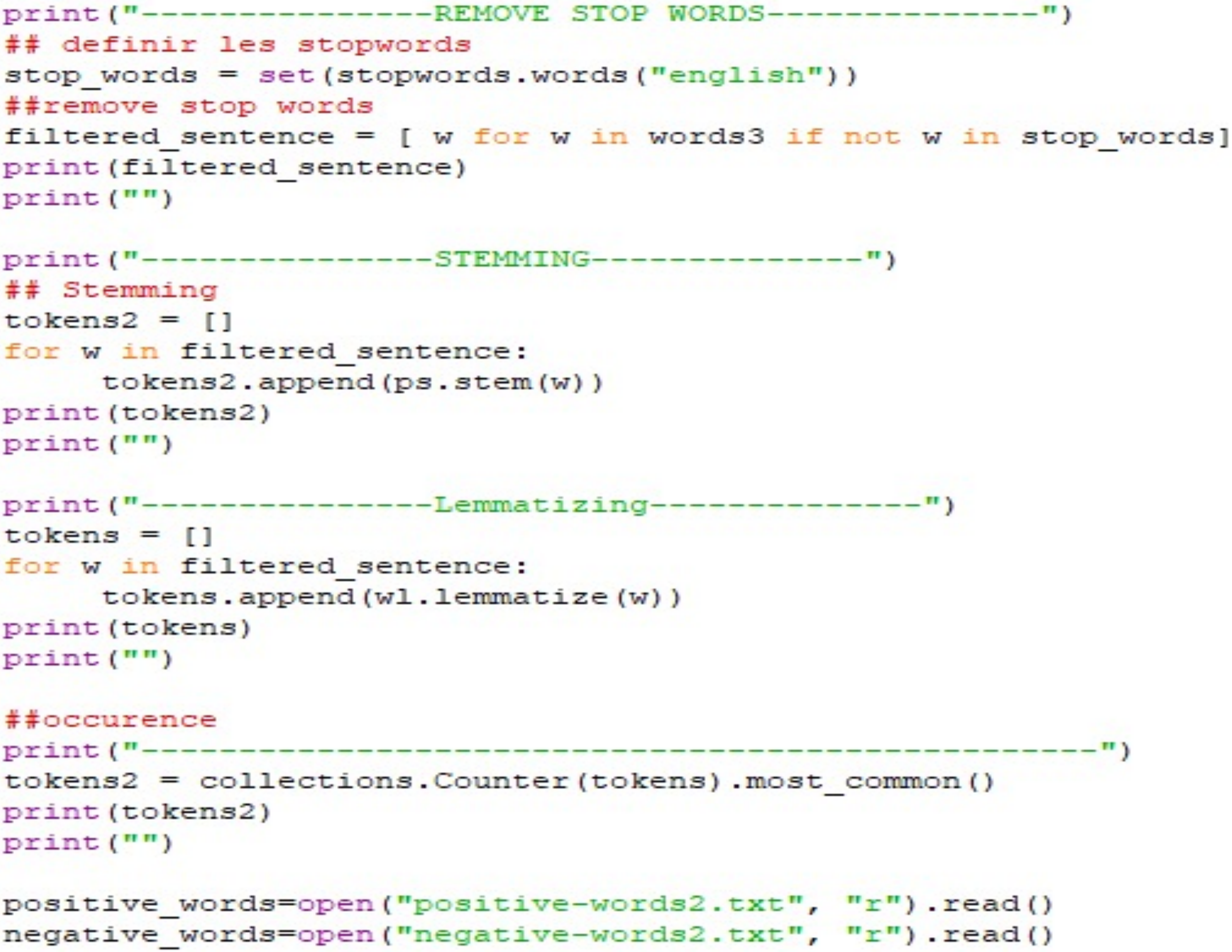
```
---------------TEXT--------------
this product is nice. i really appreciate these awsome products!

---------------TOKENIZATION AND LOWER CASE--------------
['this', 'product', 'is', 'nice', '.', 'i', 'really', 'appreciate', 'these', 'awsome', 'products', '!']

---------------NORMALIZATION--------------
['this', 'product', 'is', 'nice', 'i', 'really', 'appreciate', 'these', 'awsome', 'products']

---------------REMOVE STOP WORDS--------------
['product', 'nice', 'really', 'appreciate', 'awsome', 'products']

---------------STEMMING--------------
['product', 'nice', 'realli', 'appreci', 'awsom', 'product']

---------------Lemmatizing--------------
['product', 'nice', 'really', 'appreciate', 'awsome', 'product']

-------------------------------------------------
[('product', 2), ('nice', 1), ('really', 1), ('appreciate', 1), ('awsome', 1)]

---------------Number of positive words--------------
product
nice
appreciate
awsome
product
5

---------------Number of Negative words--------------
product
product
2

---------------Calculating percentages--------------
Positive: 83%  Negative: 33%

---------------Deciding if it is postive or negative--------------
Positive
```



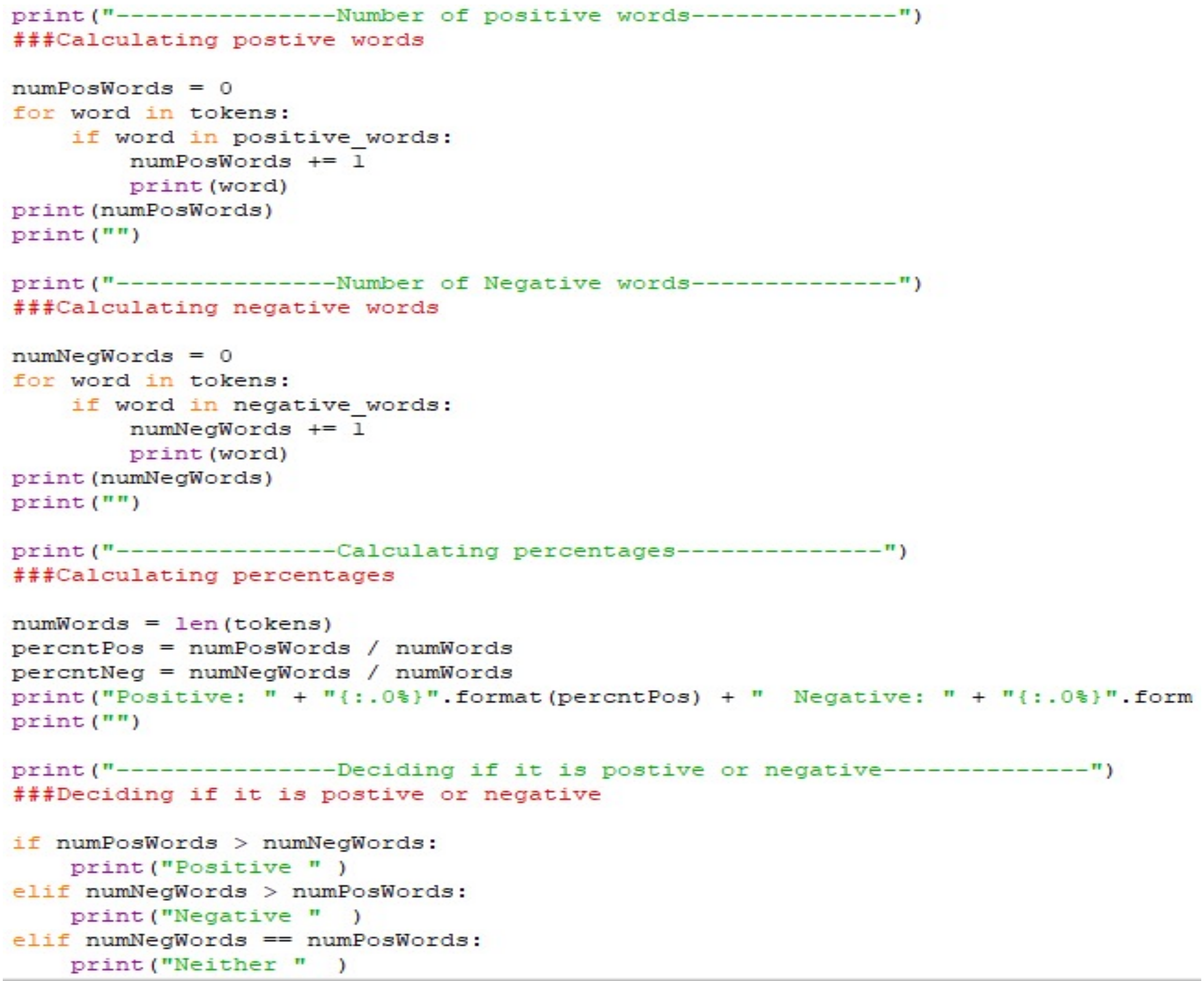
```
---------------TEXT--------------
it is a BAD and HORRIBLE movie!

---------------TOKENIZATION AND LOWER CASE--------------
['it', 'is', 'a', 'bad', 'and', 'horrible', 'movie', '!']

---------------NORMALIZATION--------------
['it', 'is', 'a', 'bad', 'and', 'horrible', 'movie']

---------------REMOVE STOP WORDS--------------
['bad', 'horrible', 'movie']

---------------STEMMING--------------
['bad', 'horribl', 'movi']

---------------Lemmatizing--------------
['bad', 'horrible', 'movie']

-------------------------------------------------
[('bad', 1), ('horrible', 1), ('movie', 1)]

---------------Number of positive words--------------
0

---------------Number of Negative words--------------
bad
horrible
2

---------------Calculating percentages--------------
Positive: 0%  Negative: 67%

---------------Deciding if it is postive or negative--------------
Negative
```



```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from nltk import tokenize
import string
import re
import collections
import sys

ps = PorterStemmer()
wl = WordNetLemmatizer()

##new_text = "It was one of the worst movies, 56 - ? despite good . \
## the movie was bad. horses, eating!"
new_text = "it is a BAD and HORRIBLE movie!"
##new_text = "this product is nice. i really appreciate these awsome products!"

print("----------------TEXT---------------")
print(new_text)
print("")

print("----------------TOKENIZATION AND LOWER CASE---------------")
## to lower case
new_text2 = new_text.lower()
##couper la phrase en mots
words = word_tokenize(new_text2)
print(words)
print("")

print("----------------NORMALIZATION---------------")
## normalisation
words2 = [x for x in words if not re.fullmatch('[' + string.punctuation + ']+',
## remove numbers
words3 = filter(lambda x: x.isalpha(), words2)
print(words2)
print("")
```

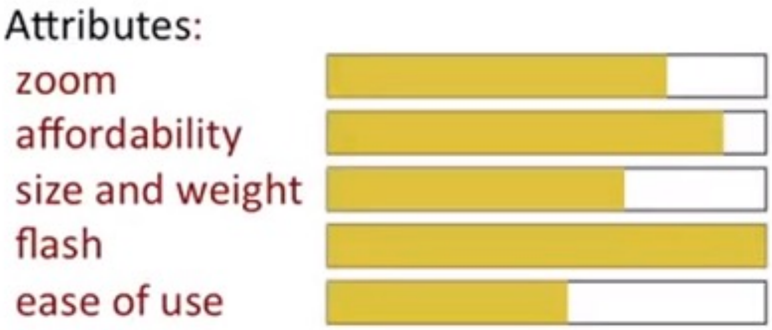


```
print("---------------REMOVE STOP WORDS--------------")
## definir les stopwords
stop_words = set(stopwords.words("english"))
##remove stop words
filtered_sentence = [ w for w in words3 if not w in stop_words]
print(filtered_sentence)
print("")

print("---------------STEMMING--------------")
## Stemming
tokens2 = []
for w in filtered_sentence:
     tokens2.append(ps.stem(w))
print(tokens2)
print("")

print("---------------Lemmatizing--------------")
tokens = []
for w in filtered_sentence:
     tokens.append(wl.lemmatize(w))
print(tokens)
print("")

##occurence
print("-------------------------------------------------")
tokens2 = collections.Counter(tokens).most_common()
print(tokens2)
print("")

positive_words=open("positive-words2.txt", "r").read()
negative_words=open("negative-words2.txt", "r").read()
```

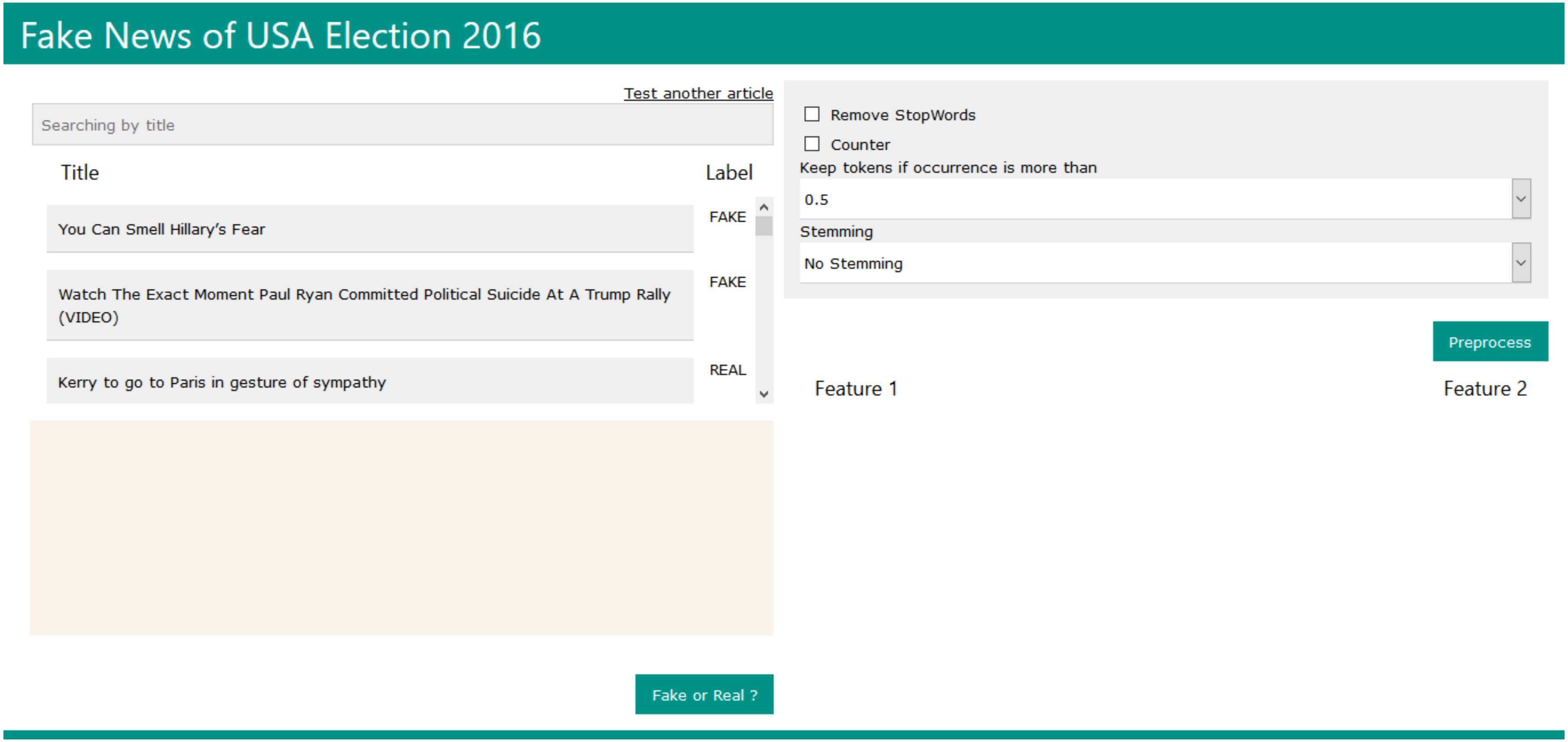


```
print("---------------Number of positive words--------------")
###Calculating postive words

numPosWords = 0
for word in tokens:
    if word in positive_words:
        numPosWords += 1
        print(word)
print(numPosWords)
print("")

print("---------------Number of Negative words--------------")
###Calculating negative words

numNegWords = 0
for word in tokens:
    if word in negative_words:
        numNegWords += 1
        print(word)
print(numNegWords)
print("")

print("---------------Calculating percentages--------------")
###Calculating percentages

numWords = len(tokens)
percntPos = numPosWords / numWords
percntNeg = numNegWords / numWords
print("Positive: " + "{:.0%}".format(percntPos) + "  Negative: " + "{:.0%}".form
print("")

print("---------------Deciding if it is postive or negative--------------")
###Deciding if it is postive or negative

if numPosWords > numNegWords:
    print("Positive " )
elif numNegWords > numPosWords:
    print("Negative "  )
elif numNegWords == numPosWords:
    print("Neither "  )
```

# Opinion mining



# *fake news*



Fake News of USA Election 2016

Test another article

Searching by title

| Title | Label |
|---|---|
| You Can Smell Hillary's Fear | FAKE |
| Watch The Exact Moment Paul Ryan Committed Political Suicide At A Trump Rally (VIDEO) | FAKE |
| Kerry to go to Paris in gesture of sympathy | REAL |

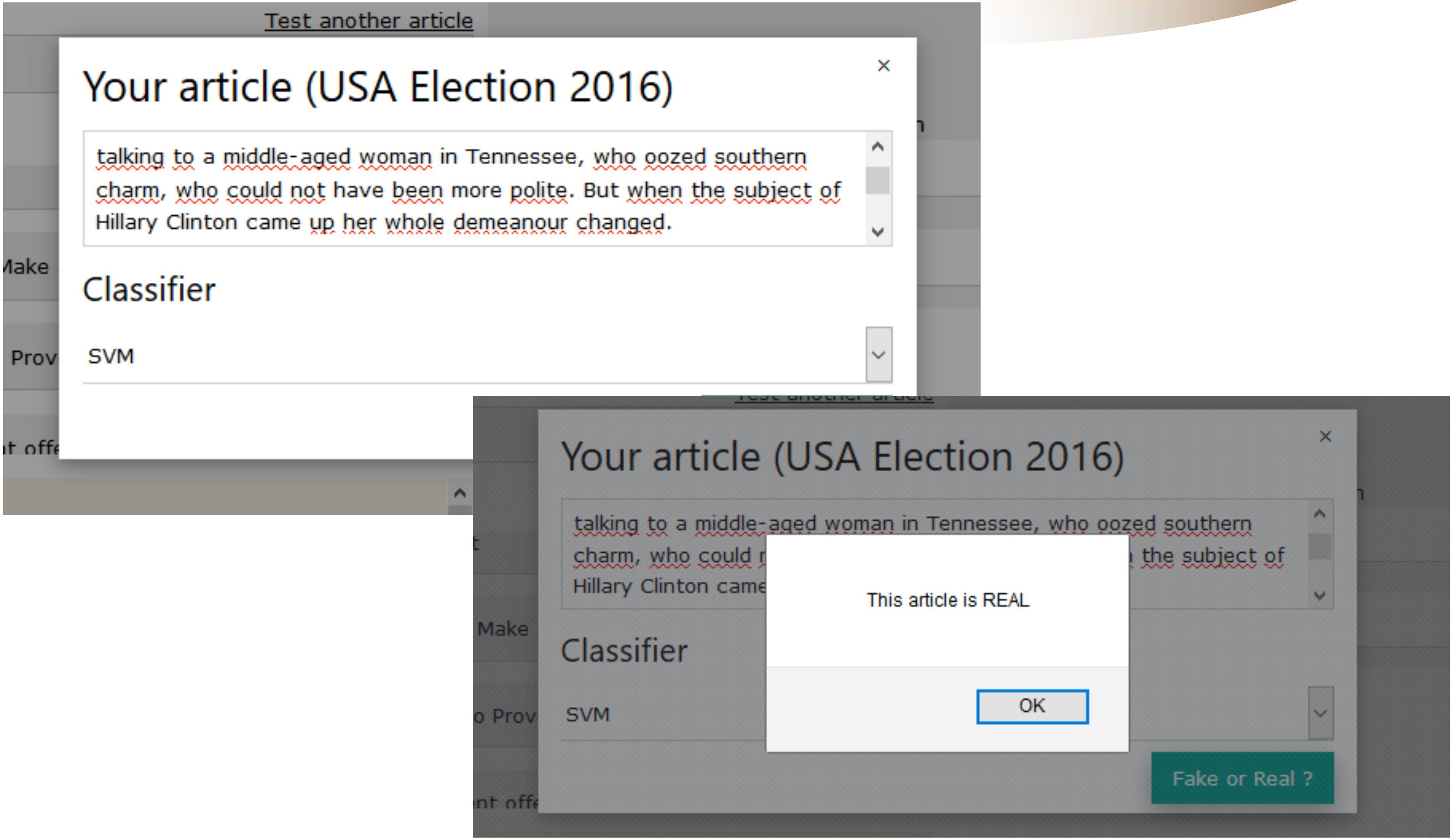
☐ Remove StopWords

☐ Counter

Keep tokens if occurrence is more than: 0.5

Stemming: No Stemming

# *fake news*



## Your article (USA Election 2016)

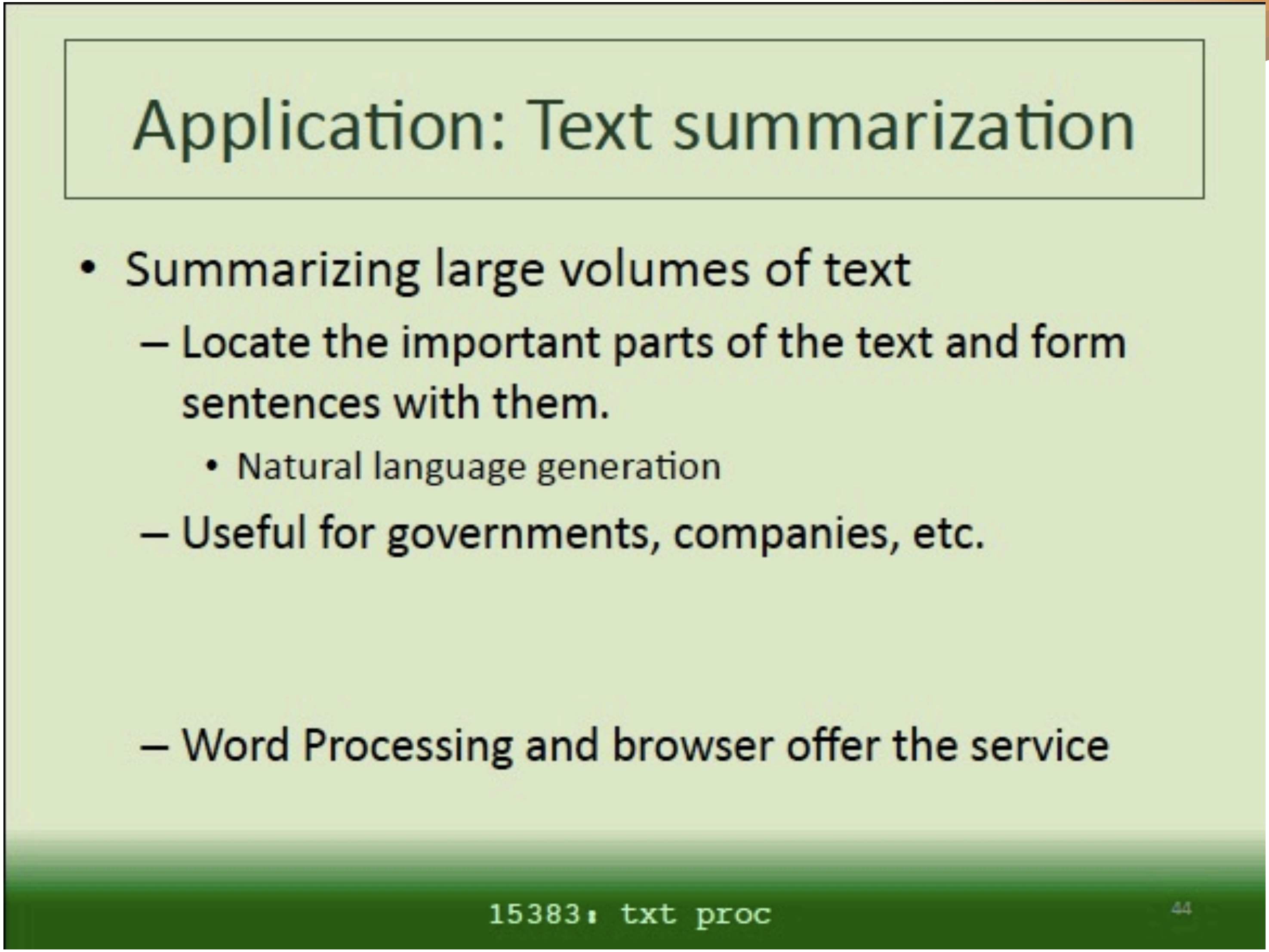
talking to a middle-aged woman in Tennessee, who oozed southern charm, who could not have been more polite. But when the subject of Hillary Clinton came up her whole demeanour changed.

### Classifier

SVM

This article is REAL

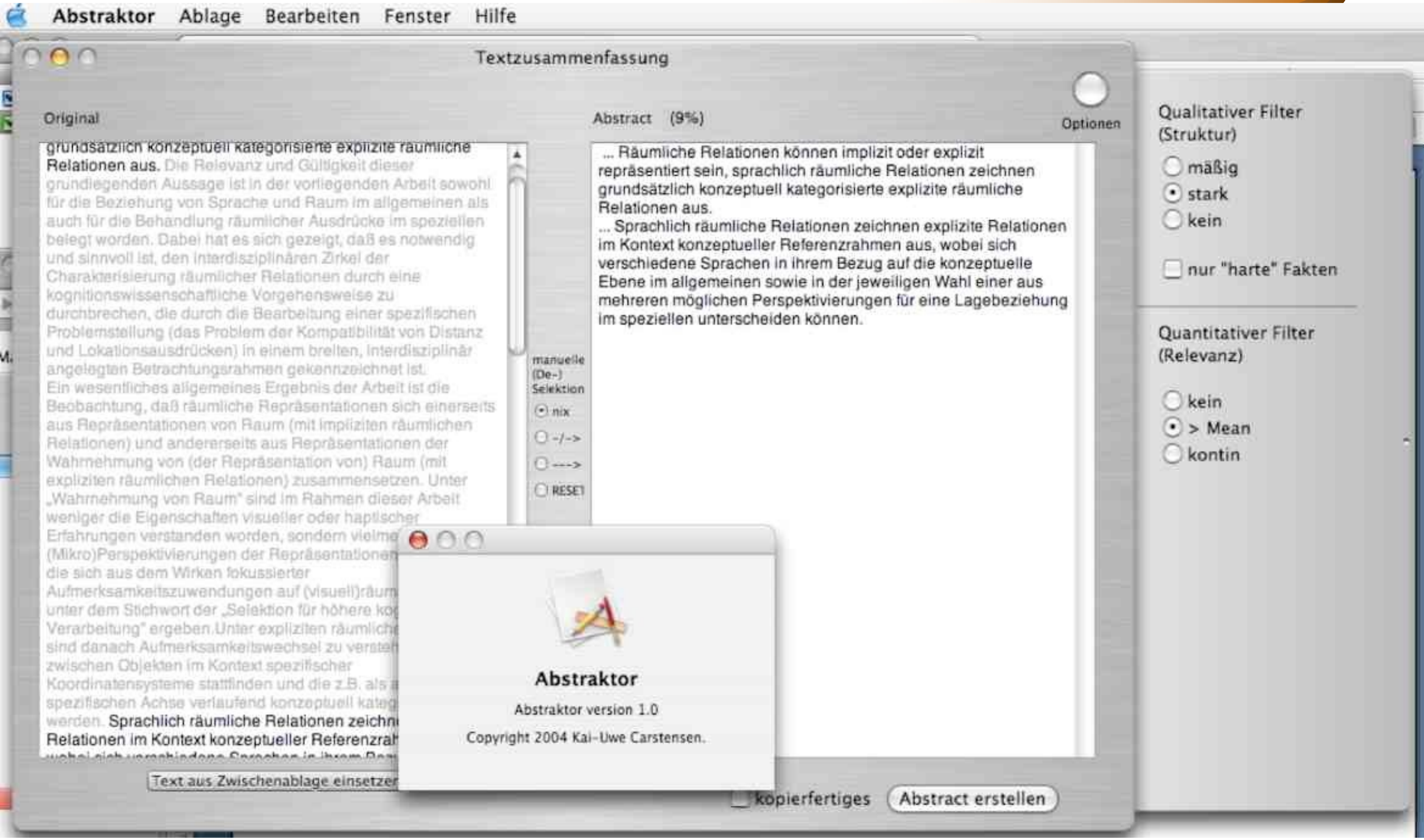
# Applications



## Application: Text summarization

- Summarizing large volumes of text
  - Locate the important parts of the text and form sentences with them.
    - Natural language generation
  - Useful for governments, companies, etc.
  - Word Processing and browser offer the service

# summarization



grundsätzlich konzeptuell kategorisierte explizite räumliche
Relationen aus. Die Relevanz und Gültigkeit dieser
grundlegenden Aussage ist in der vorliegenden Arbeit sowohl
für die Beziehung von Sprache und Raum im allgemeinen als
auch für die Behandlung räumlicher Ausdrücke im speziellen
belegt worden. Dabei hat es sich gezeigt, daß es notwendig
und sinnvoll ist, den interdisziplinären Zirkel der
Charakterisierung räumlicher Relationen durch eine
kognitionswissenschaftliche Vorgehensweise zu
durchbrechen, die durch die Bearbeitung einer spezifischen
Problemstellung (das Problem der Kompatibilität von Distanz
und Lokationsausdrücken) in einem breiten, interdisziplinär
angelegten Betrachtungsrahmen gekennzeichnet ist.
Ein wesentliches allgemeines Ergebnis der Arbeit ist die
Beobachtung, daß räumliche Repräsentationen sich einerseits
aus Repräsentationen von Raum (mit impliziten räumlichen
Relationen) und andererseits aus Repräsentationen der
Wahrnehmung von (der Repräsentation von) Raum (mit
expliziten räumlichen Relationen) zusammensetzen. Unter
„Wahrnehmung von Raum" sind im Rahmen dieser Arbeit
weniger die Eigenschaften visueller oder haptischer
Erfahrungen verstanden worden, sondern vielme
(Mikro)Perspektivierungen der Repräsentatione
die sich aus dem Wirken fokussierter
Aufmerksamkeitszuwendungen auf (visuell)räum
unter dem Stichwort der „Selektion für höhere ko
Verarbeitung" ergeben.Unter expliziten räumlich
sind danach Aufmerksamkeitswechsel zu verste
zwischen Objekten im Kontext spezifischer
Koordinatensysteme stattfinden und die z.B. als
spezifischen Achse verlaufend konzeptuell kateg
werden. Sprachlich räumliche Relationen zeichn
Relationen im Kontext konzeptueller Referenzra

... Räumliche Relationen können implizit oder explizit repräsentiert sein, sprachlich räumliche Relationen zeichnen grundsätzlich konzeptuell kategorisierte explizite räumliche Relationen aus.
... Sprachlich räumliche Relationen zeichnen explizite Relationen im Kontext konzeptueller Referenzrahmen aus, wobei sich verschiedene Sprachen in ihrem Bezug auf die konzeptuelle Ebene im allgemeinen sowie in der jeweiligen Wahl einer aus mehreren möglichen Perspektivierungen für eine Lagebeziehung im speziellen unterscheiden können.

manuelle (De-) Selektion
- ● nix
- ○ -/->
- ○ --->
- ○ RESET

Qualitativer Filter (Struktur)
- ○ mäßig
- ● stark
- ○ kein

□ nur "harte" Fakten

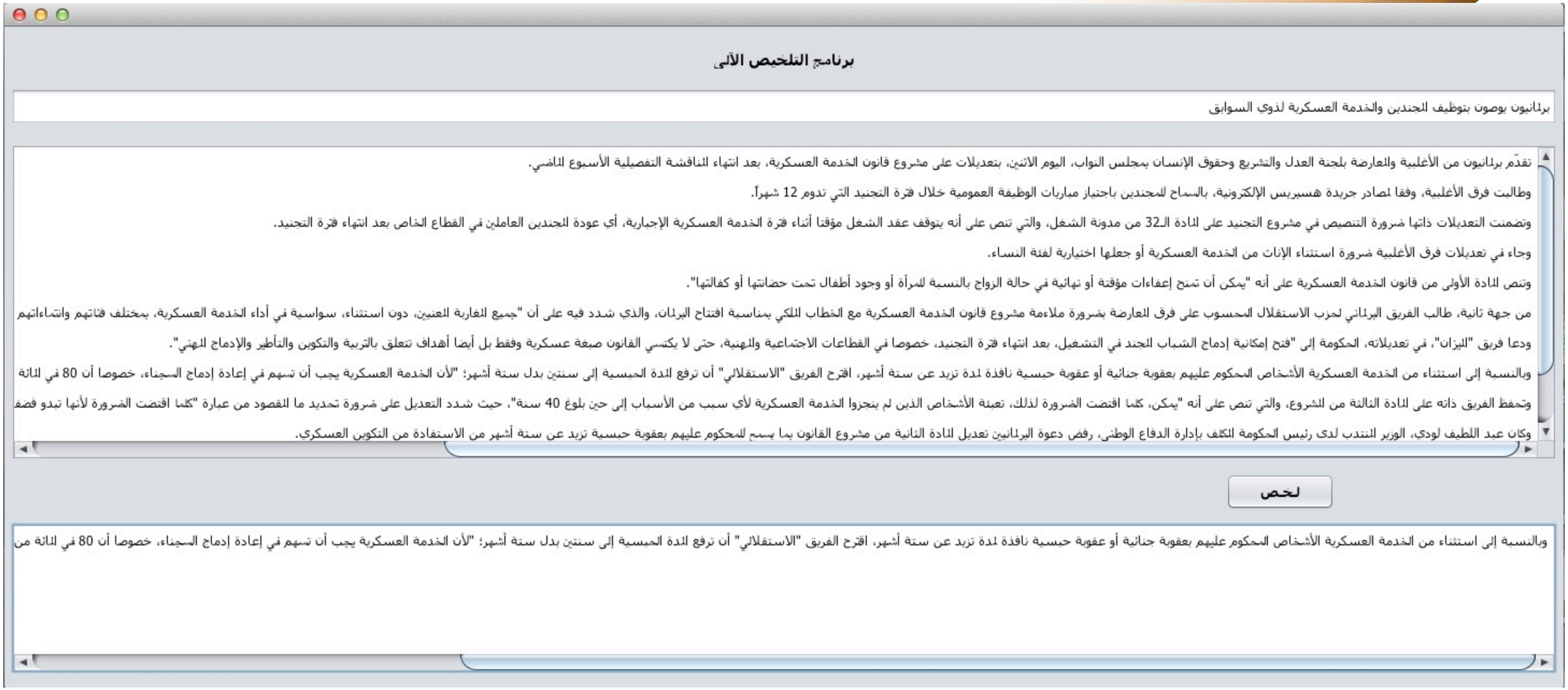
Quantitativer Filter (Relevanz)
- ○ kein
- ● > Mean
- ○ kontin

Abstraktor

Abstraktor version 1.0

Copyright 2004 Kai-Uwe Carstensen.

# appl / services – summarization



برلمانيون يوصون بتوظيف المجندين والخدمة العسكرية لذوي السوابق

تقدّم برلمانيون من الأغلبية والمعارضة بلجنة العدل والتشريع وحقوق الإنسان بمجلس النواب، اليوم الاثنين، بتعديلات على مشروع قانون الخدمة العسكرية، بعد انتهاء المناقشة التفصيلية الأسبوع الماضي.

وطالبت فرق الأغلبية، وفقا لمصادر جريدة هسبريس الإلكترونية، بالسماح للمجندين باجتياز مباريات الوظيفة العمومية خلال فترة التجنيد التي تدوم 12 شهراً.

وتضمنت التعديلات ذاتها ضرورة التنصيص في مشروع التجنيد على المادة الـ32 من مدونة الشغل، والتي تنص على أنه يتوقف عقد الشغل مؤقتا أثناء فترة الخدمة العسكرية الإجبارية، أي عودة المجندين العاملين في القطاع الخاص بعد انتهاء فترة التجنيد.

وجاء في تعديلات فرق الأغلبية ضرورة استثناء الإناث من الخدمة العسكرية أو جعلها اختيارية لفئة النساء.

وتنص المادة الأولى من قانون الخدمة العسكرية على أنه "يمكن أن تمنح إعفاءات مؤقتة أو نهائية في حالة الزواج بالنسبة للمرأة أو وجود أطفال تحت حضانتها أو كفالتها".

من جهة ثانية، طالب الفريق البرلماني لحزب الاستقلال المحسوب على فرق المعارضة بضرورة ملاءمة مشروع قانون الخدمة العسكرية مع الخطاب الملكي بمناسبة افتتاح البرلمان، والذي شدد فيه على أن "جميع المغاربة المعنيين، دون استثناء، سواسية في أداء الخدمة العسكرية، بمختلف فئاتهم وانتماءاتهم

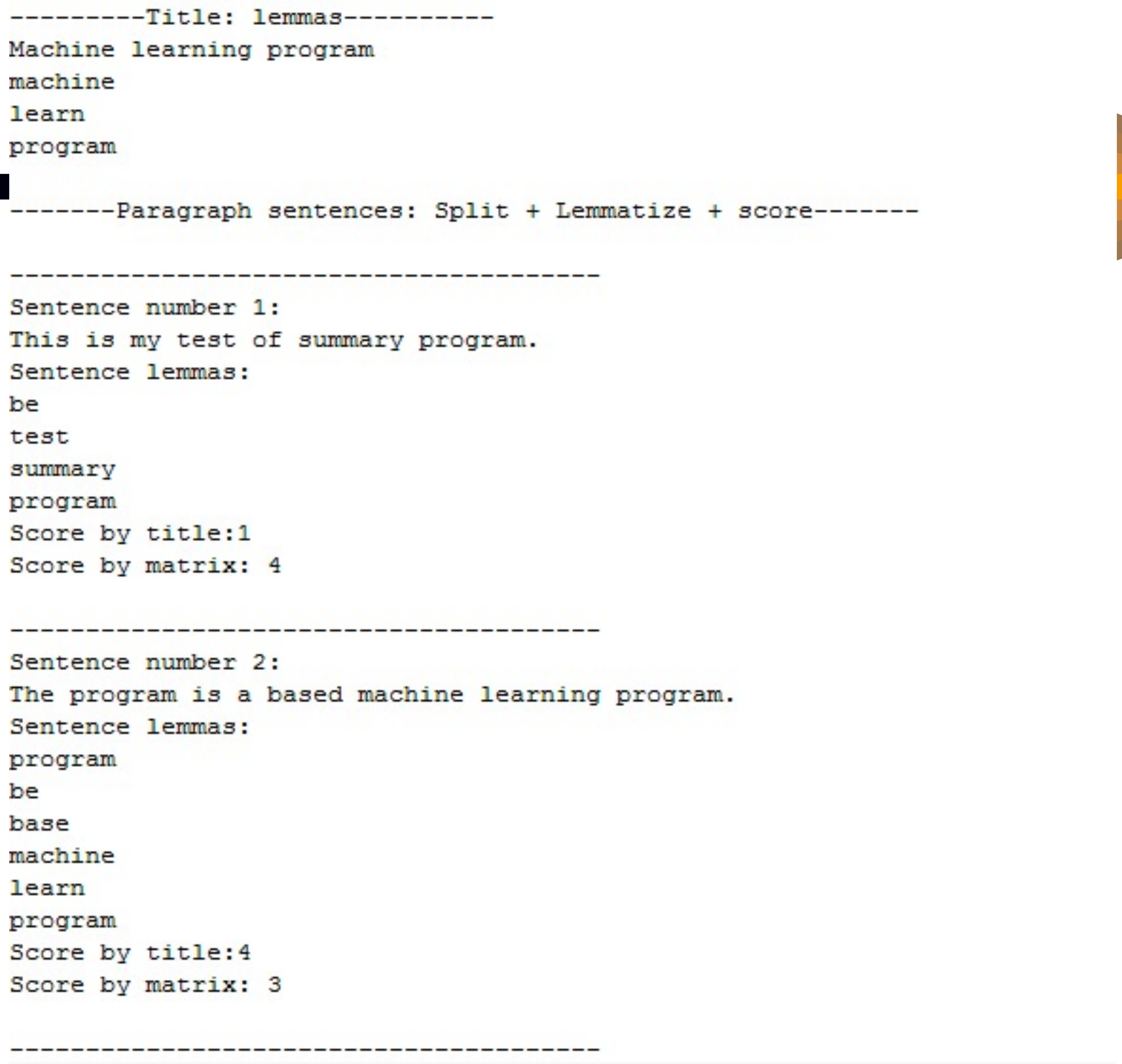
ودعا فريق "الميزان"، في تعديلاته، الحكومة إلى "فتح إمكانية إدماج الشباب المجند في التشغيل، بعد انتهاء فترة التجنيد، خصوصا في القطاعات الاجتماعية والمهنية، حتى لا يكتسي القانون صبغة عسكرية وفقط بل أيضا أهداف تتعلق بالتربية والتكوين والتأطير والإدماج المهني".

وبالنسبة إلى استثناء من الخدمة العسكرية الأشخاص المحكوم عليهم بعقوبة جنائية أو عقوبة حبسية نافذة لمدة تزيد عن ستة أشهر، اقترح الفريق "الاستقلالي" أن ترفع المدة الحبسية إلى سنتين بدل ستة أشهر؛ "لأن الخدمة العسكرية يجب أن تسهم في إعادة إدماج السجناء، خصوصا أن 80 في المائة

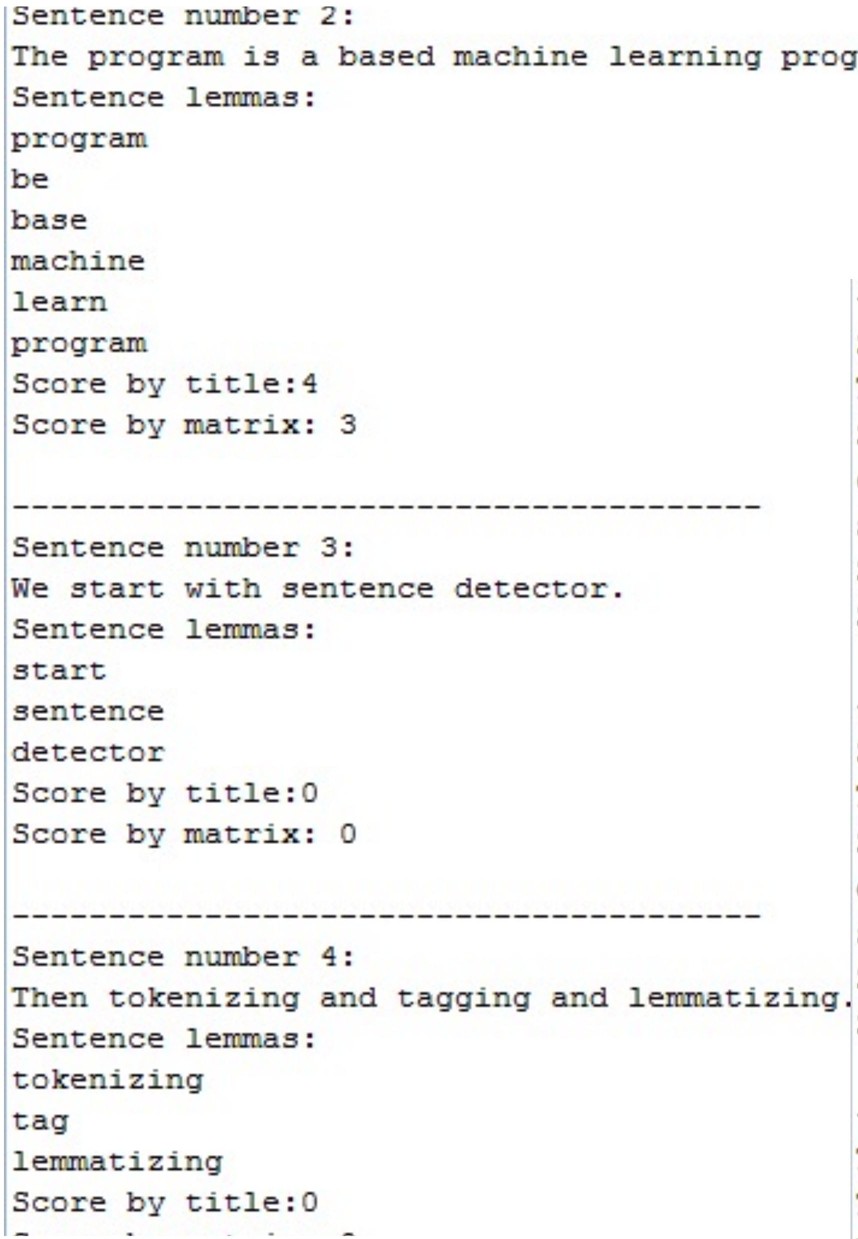
وتحفظ الفريق ذاته على المادة الثالثة من المشروع، والتي تنص على أنه "يمكن، كلما اقتضت الضرورة لذلك، تعبئة الأشخاص الذين لم ينجزوا الخدمة العسكرية لأي سبب من الأسباب إلى حين بلوغ 40 سنة"، حيث شدد التعديل على ضرورة تحديد ما المقصود من عبارة "كلما اقتضت الضرورة لأنها تبدو فضف

وكان عبد اللطيف لودي، الوزير المنتدب لدى رئيس الحكومة المكلف بإدارة الدفاع الوطني، رفض دعوة البرلمانيين تعديل المادة الثانية من مشروع القانون بما يسمح للمحكوم عليهم بعقوبة حبسية تزيد عن ستة أشهر من الاستفادة من التكوين العسكري.

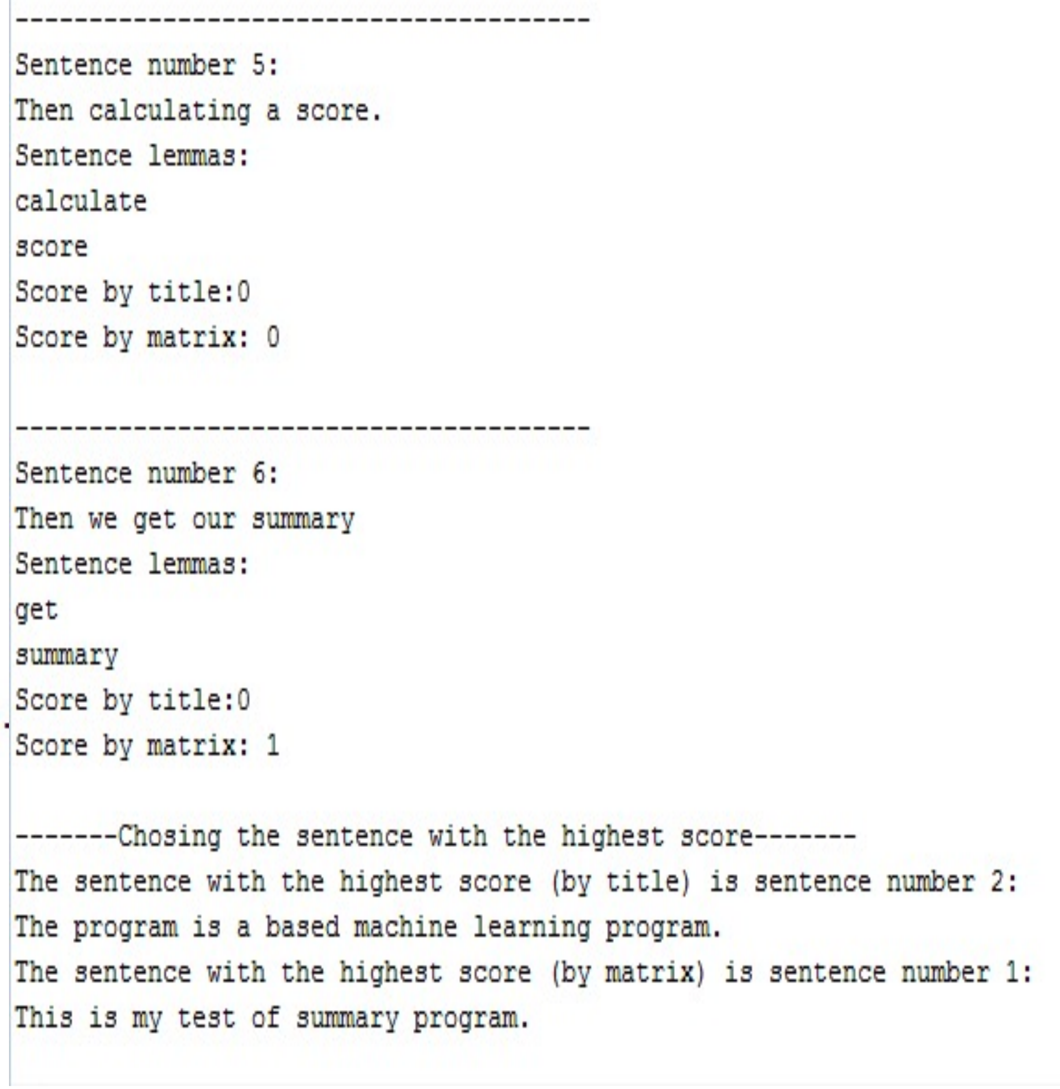
وبالنسبة إلى استثناء من الخدمة العسكرية الأشخاص المحكوم عليهم بعقوبة جنائية أو عقوبة حبسية نافذة لمدة تزيد عن ستة أشهر، اقترح الفريق "الاستقلالي" أن ترفع المدة الحبسية إلى سنتين بدل ستة أشهر؛ "لأن الخدمة العسكرية يجب أن تسهم في إعادة إدماج السجناء، خصوصا أن 80 في المائة من



```
---------Title: lemmas----------
Machine learning program
machine
learn
program

-------Paragraph sentences: Split + Lemmatize + score-------

---------------------------------------
Sentence number 1:
This is my test of summary program.
Sentence lemmas:
be
test
summary
program
Score by title:1
Score by matrix: 4

---------------------------------------
Sentence number 2:
The program is a based machine learning program.
Sentence lemmas:
program
be
base
machine
learn
program
Score by title:4
Score by matrix: 3

---------------------------------------
```



```
Sentence number 2:
The program is a based machine learning program.
Sentence lemmas:
program
be
base
machine
learn
program
Score by title:4
Score by matrix: 3

---------------------------------------
Sentence number 3:
We start with sentence detector.
Sentence lemmas:
start
sentence
detector
Score by title:0
Score by matrix: 0

---------------------------------------
Sentence number 4:
Then tokenizing and tagging and lemmatizing.
Sentence lemmas:
tokenizing
tag
lemmatizing
Score by title:0
```

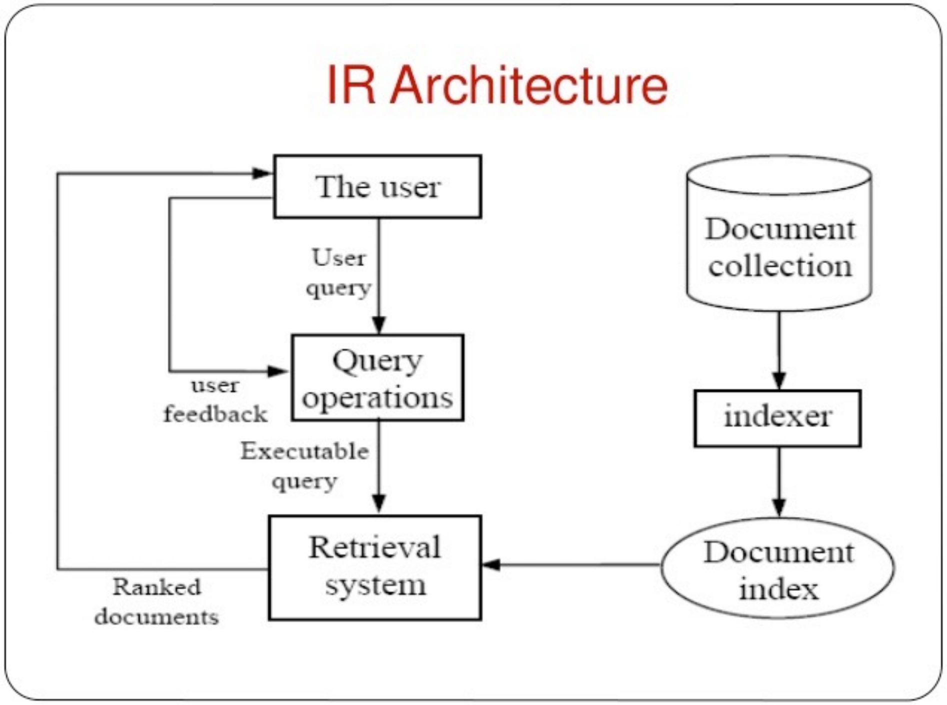


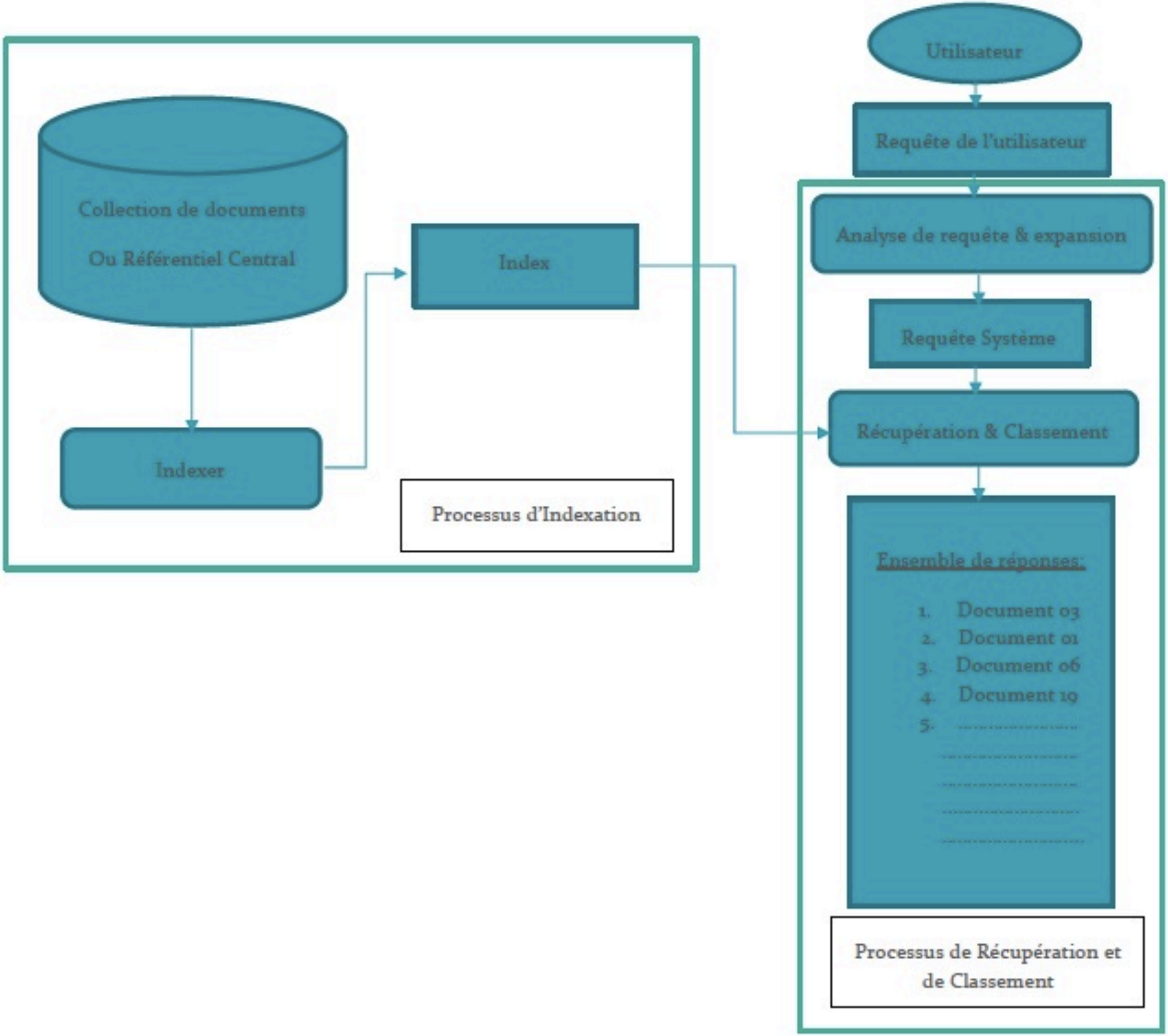
```
---------------------------------------
Sentence number 5:
Then calculating a score.
Sentence lemmas:
calculate
score
Score by title:0
Score by matrix: 0

---------------------------------------
Sentence number 6:
Then we get our summary
Sentence lemmas:
get
summary
Score by title:0
Score by matrix: 1

-------Chosing the sentence with the highest score-------
The sentence with the highest score (by title) is sentence number 2:
The program is a based machine learning program.
The sentence with the highest score (by matrix) is sentence number 1:
This is my test of summary program.
```

# Applications

IR:

- Save documents (or their addresses) and determine a set of characteristics according to their analysis
- Build accessible and regularly updated indexes
- Answer queries by selecting the most relevant documents





Collection de documents

Ou Référentiel Central

Indexer

Index

Processus d'Indexation

Utilisateur

Requête de l'utilisateur

Analyse de requête & expansion

Requête Système

Récupération & Classement

Ensemble de réponses:

1. Document 03
2. Document 01
3. Document 06
4. Document 19
5. ………………………

………………………

………………………

………………………

………………………

Processus de Récupération et de Classement

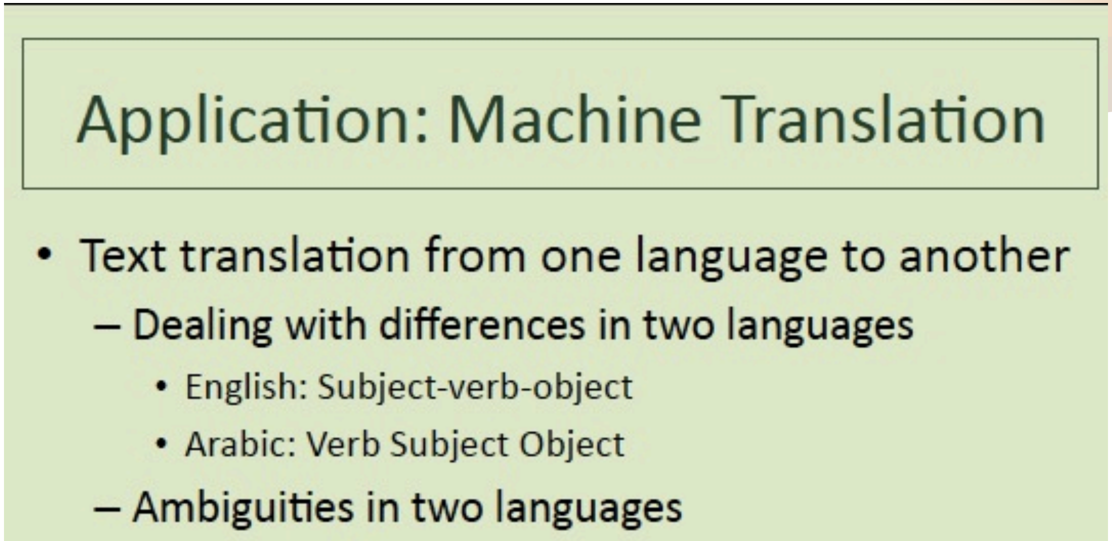
# *Applications*

## Spell checking:

- Identify words (tokenization)
- Orthographic correction: correct the words that belong to the dictionary and that are not in a foreign language, nor named entities, numbers, acronyms ...
- Grammar correction: determine the function of the words within the sentence (determinant, noun, verb, adverb, etc.) then to carry out a syntactic analysis

- http://arabic.emi.ac.ma:8080/Medictionnary/

# Applications

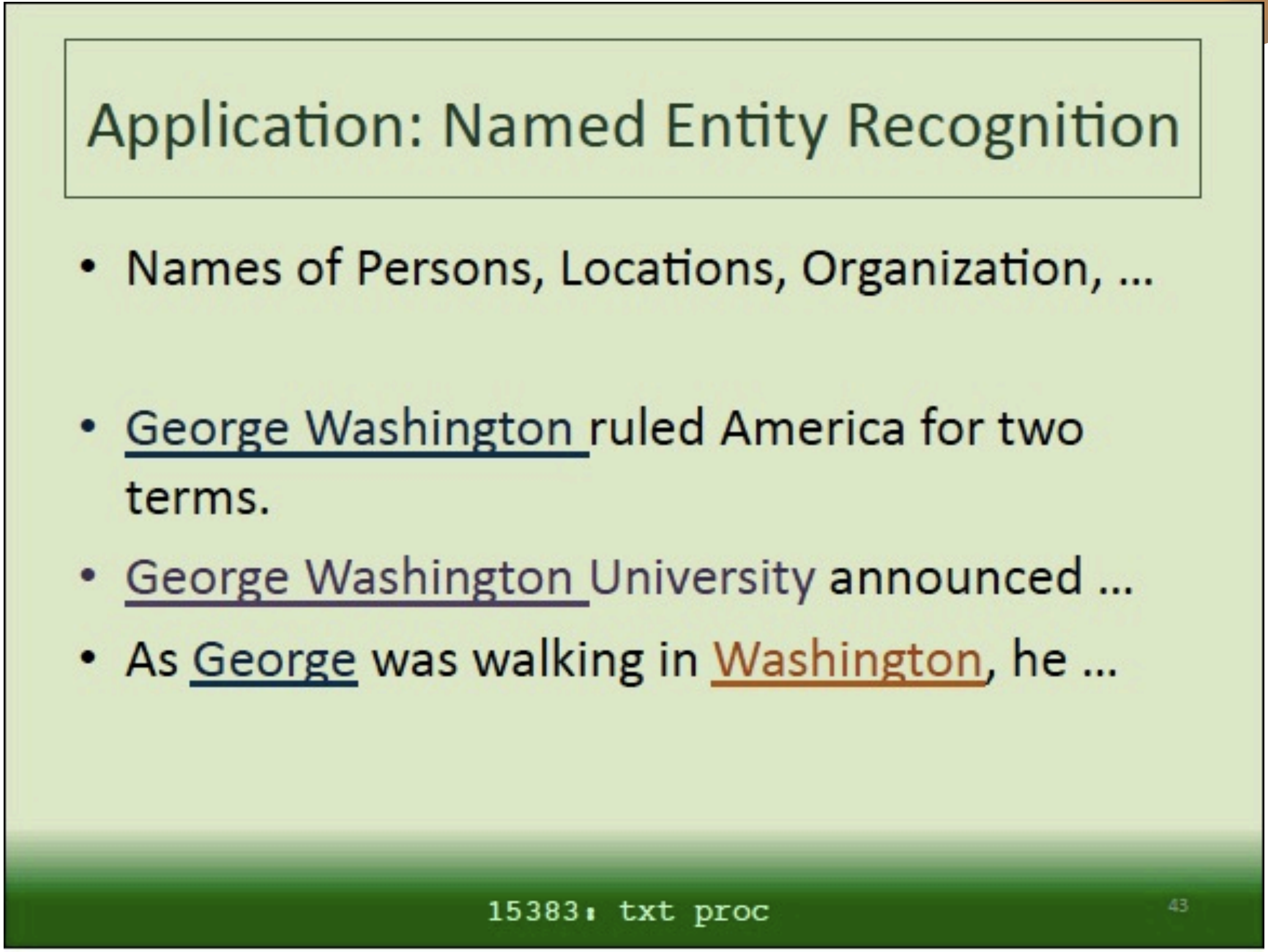


## Application: Machine Translation

- Text translation from one language to another
  - Dealing with differences in two languages
    - English: Subject-verb-object
    - Arabic: Verb Subject Object
  - Ambiguities in two languages

- Obvious application interest, but particularly difficult task
- Current quality not exceptional but sufficient to be useful
- Several online translation:
- https://www.babelfish.com/
- https://www.bing.com/translator
- http://www.reverso.net/
- https://translate.google.com/

# Applications



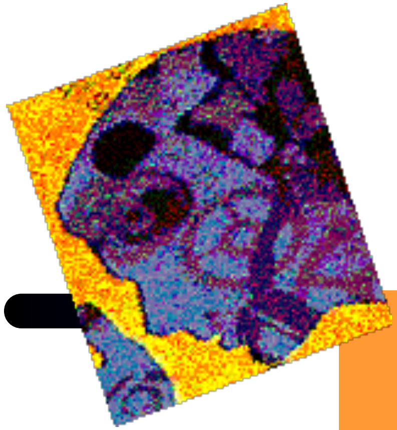
## Application: Named Entity Recognition

- Names of Persons, Locations, Organization, …
- George Washington ruled America for two terms.
- George Washington University announced …
- As George was walking in Washington, he …

# Development

- www.nltk.org
- www.gate.ac.uk
- uima.apache.org

- arabic.emi.ac.ma/safar
- camel.abudhabi.nyu.edu/madamira/



# *Natural Language Processing*
# *Computational Linguistics*
# *Text processing*