Natural Language Processing

Presented By :

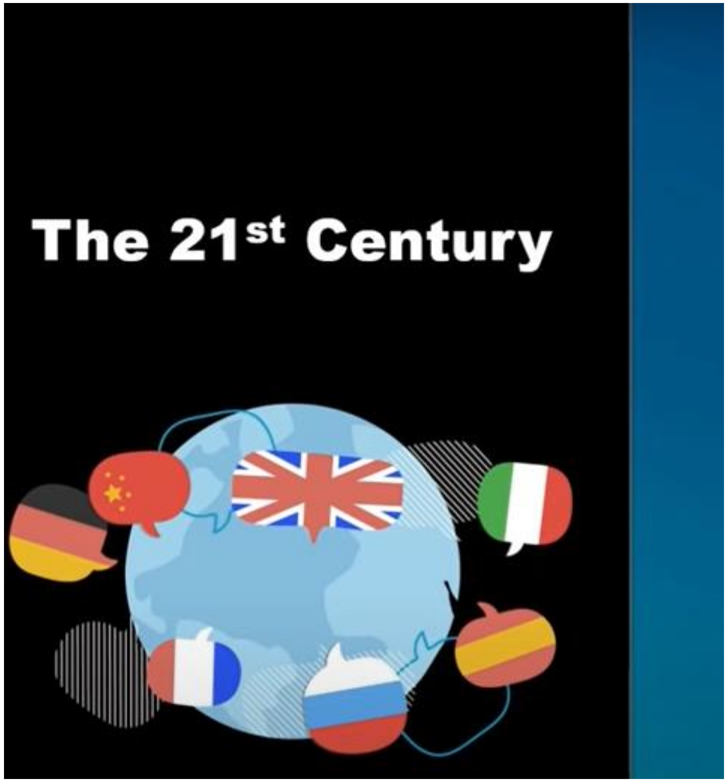
Prof.(Dr.) Archana Singh

Head – Dept of Artificial Intelligence
ASET, Amity University, Noida

LEARNING OUTCOMES

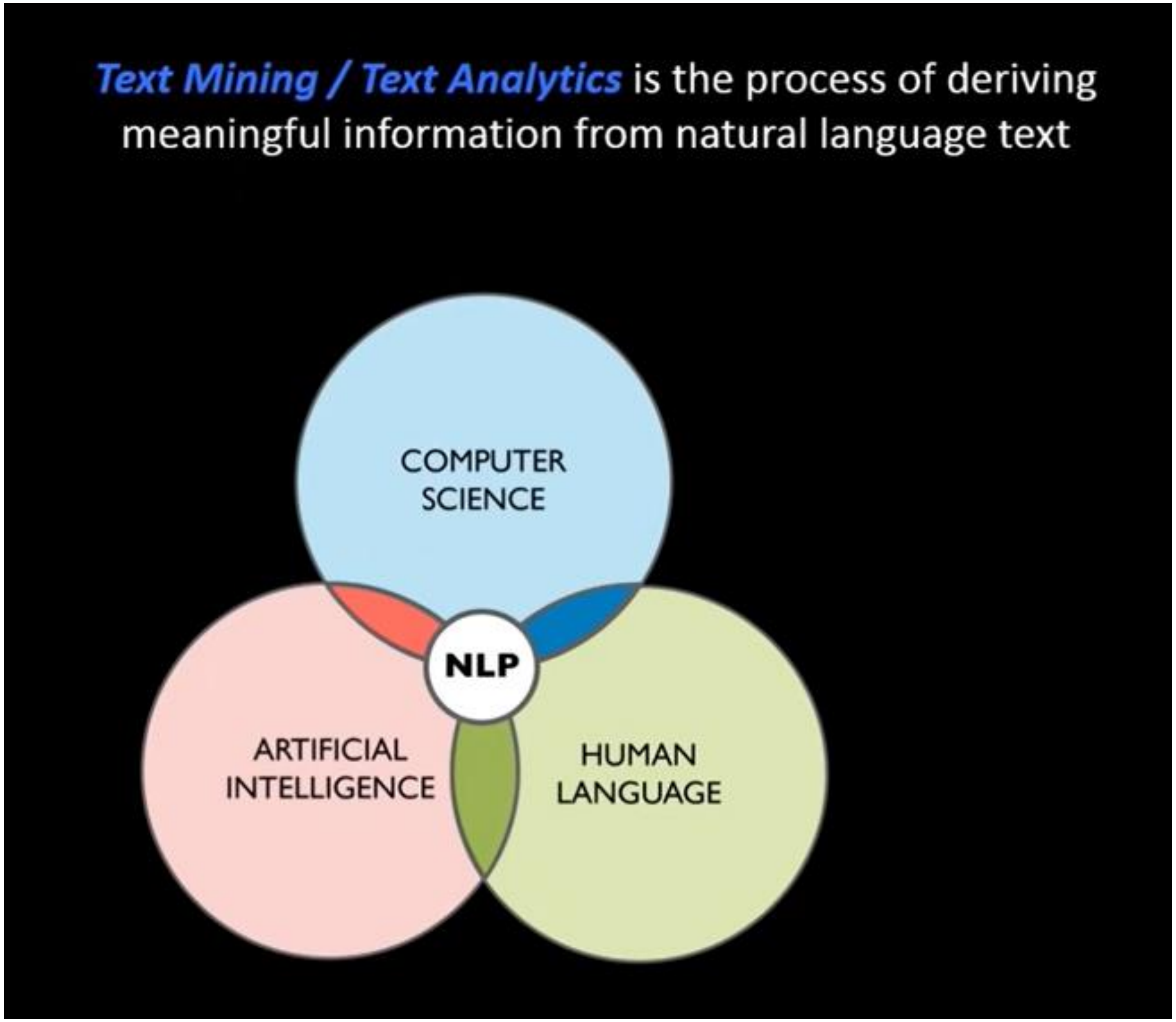
- Students will be able to know about NLP and basics of syntactic processing
- Students will be able to understand parsing techniques
- Students will be able to analyze a problem of NLP and its research applications

Natural Language Processing



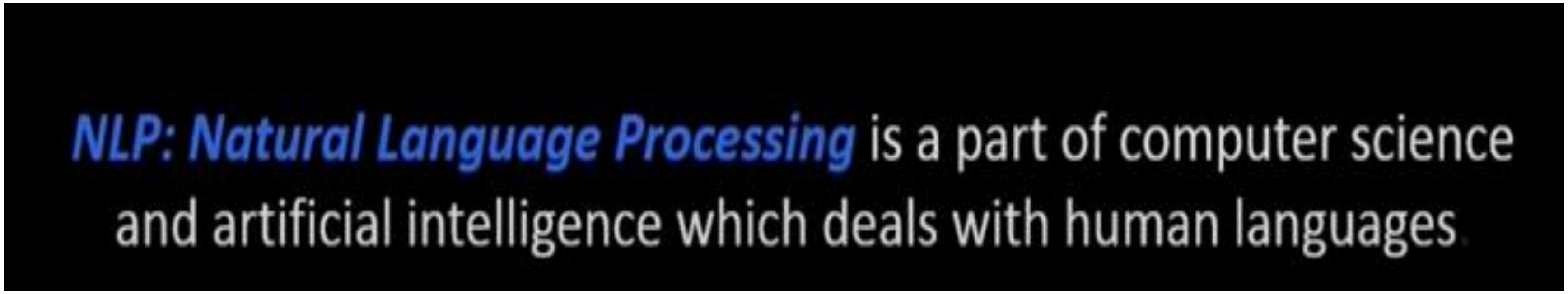
6500 Languages

Positioning of NLP



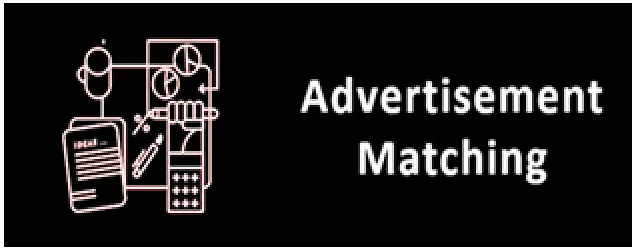
Text Mining / Text Analytics is the process of deriving meaningful information from natural language text

What is he saying


NLP: Natural Language Processing is a part of computer science and artificial intelligence which deals with human languages.

Why do we need NLP

Applications of NLP



Siri

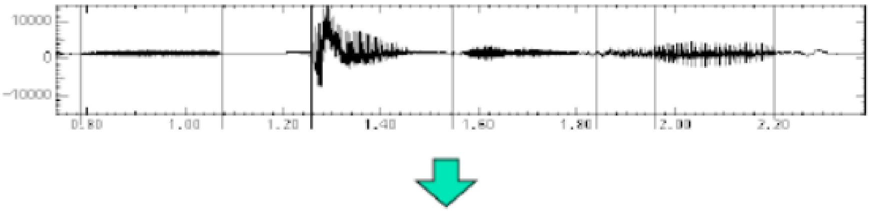
Text to Speech : Text –in Audio-out

Speech Recognition

Language Analysis

Dialog Processing

Text to Speech



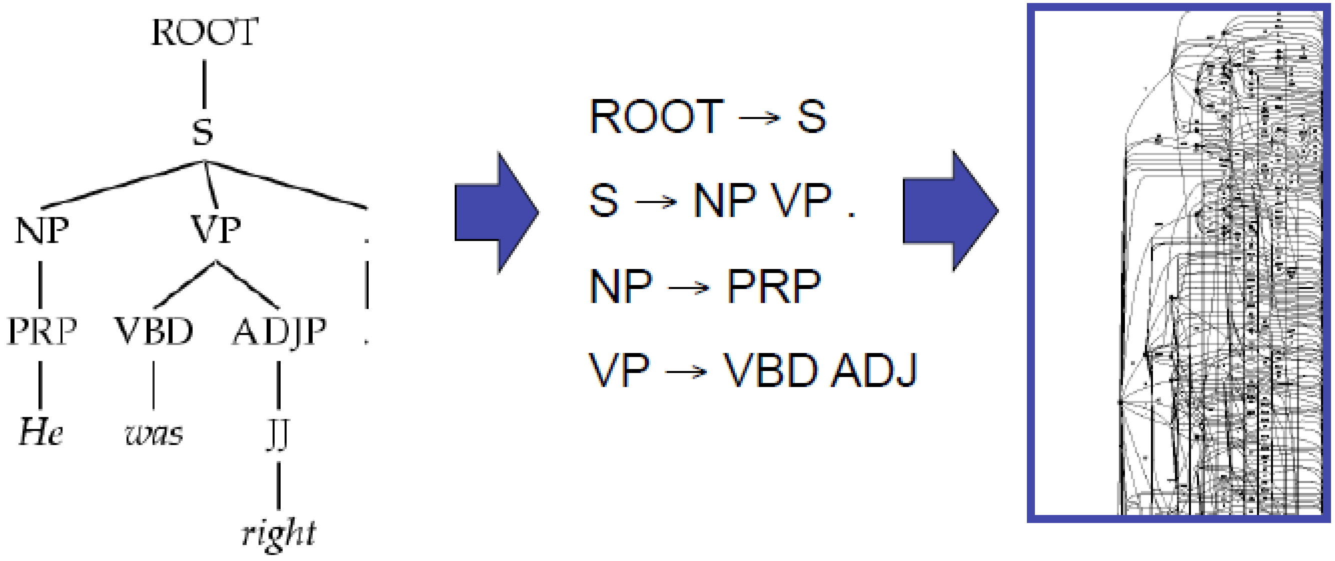
“Speech Lab”

Corpora

- A corpus is a collection of text
  - Often annotated in some way
  - Sometimes just lots of text
  - Balanced vs. uniform corpora
- Examples
  - Newswire collections: 500M+ words
  - Brown corpus: 1M words of tagged "balanced" text
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of aligned French / English sentences
  - The Web: billions of words of who knows what

Corpus based Methods

- A corpus like a treebank gives us three important tools:
  - It gives us broad coverage



ROOT → S

S → NP VP .

NP → PRP

VP → VBD ADJ

Semantic Analysis

- NLP is much more than syntax!
- Even correct tree structured syntactic analyses don’t fully nail down the meaning

I haven’t slept for ten days

John’s boss said he was doing better

- In general, every level of linguistic structure comes with its own ambiguities…

Understanding languages

- Tokenization/morphology:
  - What are the words, what is the sub-word structure?
  - Often simple rules work (period after “Mr.” isn’t sentence break)
  - Relatively easy in English, other languages are harder:
    - Segementation

      哲学家维特根斯坦出生于维也纳

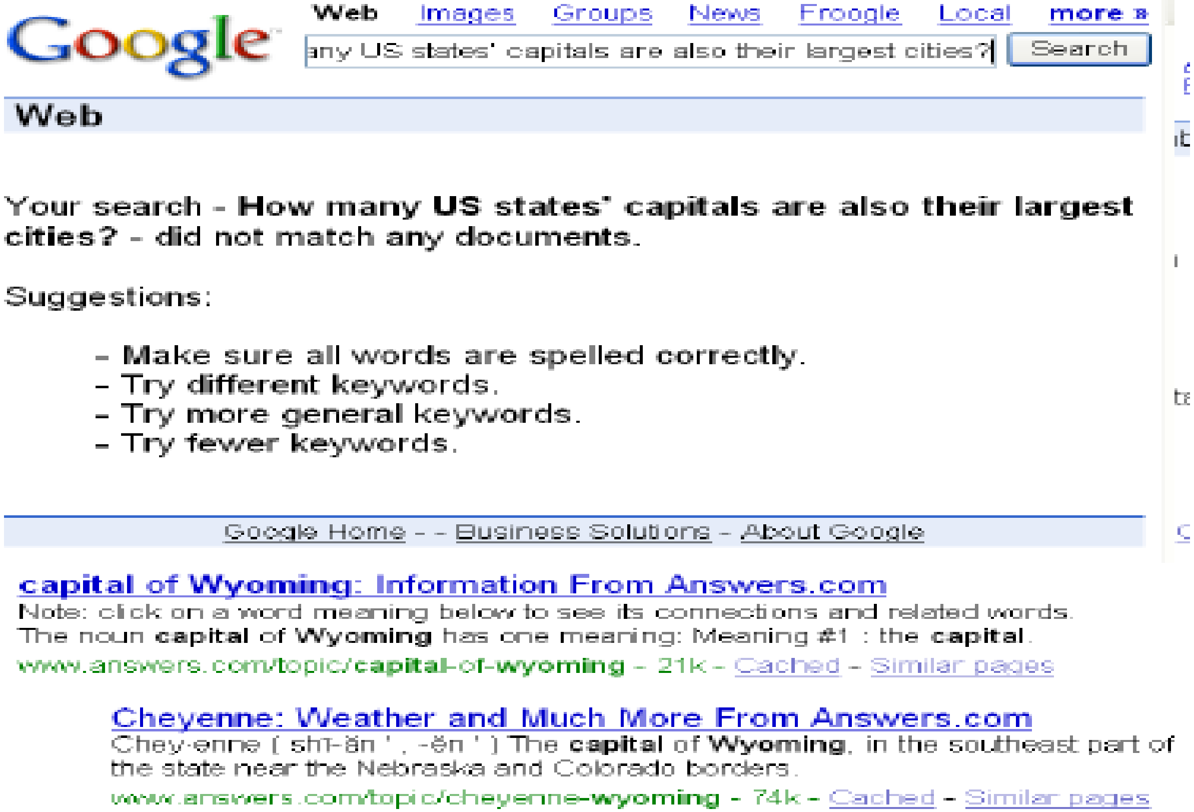
    - Morphology

sarà	*andata*
be+fut+3sg	go+ppt+fem

“she will have gone”

- Discourse: how do sentences relate to each other?
- Pragmatics: what intent is expressed by the literal meaning, how to react to an utterance?
- Phonetics: acoustics and physical production of sounds
- Phonology: how sounds pattern in a language

Information Retrieval



Your search - **How many US states' capitals are also their largest cities?** - did not match any documents.

Suggestions:

- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

Google Home - - Business Solutions - About Google

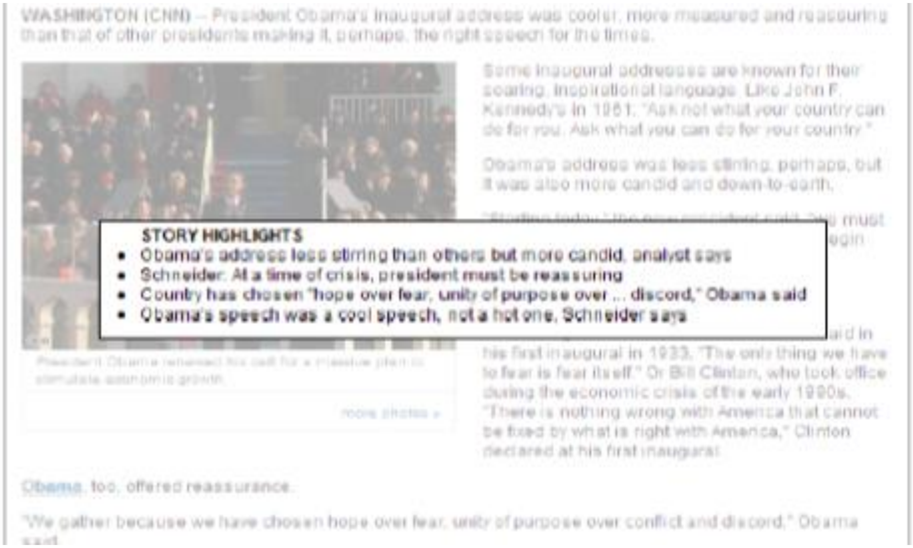
capital of Wyoming: Information From Answers.com
Note: click on a word meaning below to see its connections and related words.
The noun **capital** of **Wyoming** has one meaning: Meaning #1 : the **capital**.
www.answers.com/topic/**capital-of-wyoming** - 21k - Cached - Similar pages

Cheyenne: Weather and Much More From Answers.com
Chey·enne (shī-ăn ' , -ĕn ') The **capital** of **Wyoming**, in the southeast part of the state near the Nebraska and Colorado borders.
www.answers.com/topic/cheyenne-**wyoming** - 74k - Cached - Similar pages

Text Summarization

Considering documents

An example of analysis with text generation



WASHINGTON (CNN) -- President Obama's inaugural address was cooler, more measured and reassuring than that of other presidents making it, perhaps, the right speech for the times.

President Obama renewed his call for a massive plan to stimulate economic growth.

more photos »

Some inaugural addresses are known for their soaring, inspirational language. Like John F. Kennedy's in 1961: "Ask not what your country can do for you. Ask what you can do for your country."

Obama's address was less stirring, perhaps, but it was also more candid and down-to-earth.

"Starting today," the new president said, "we must ... egin

STORY HIGHLIGHTS

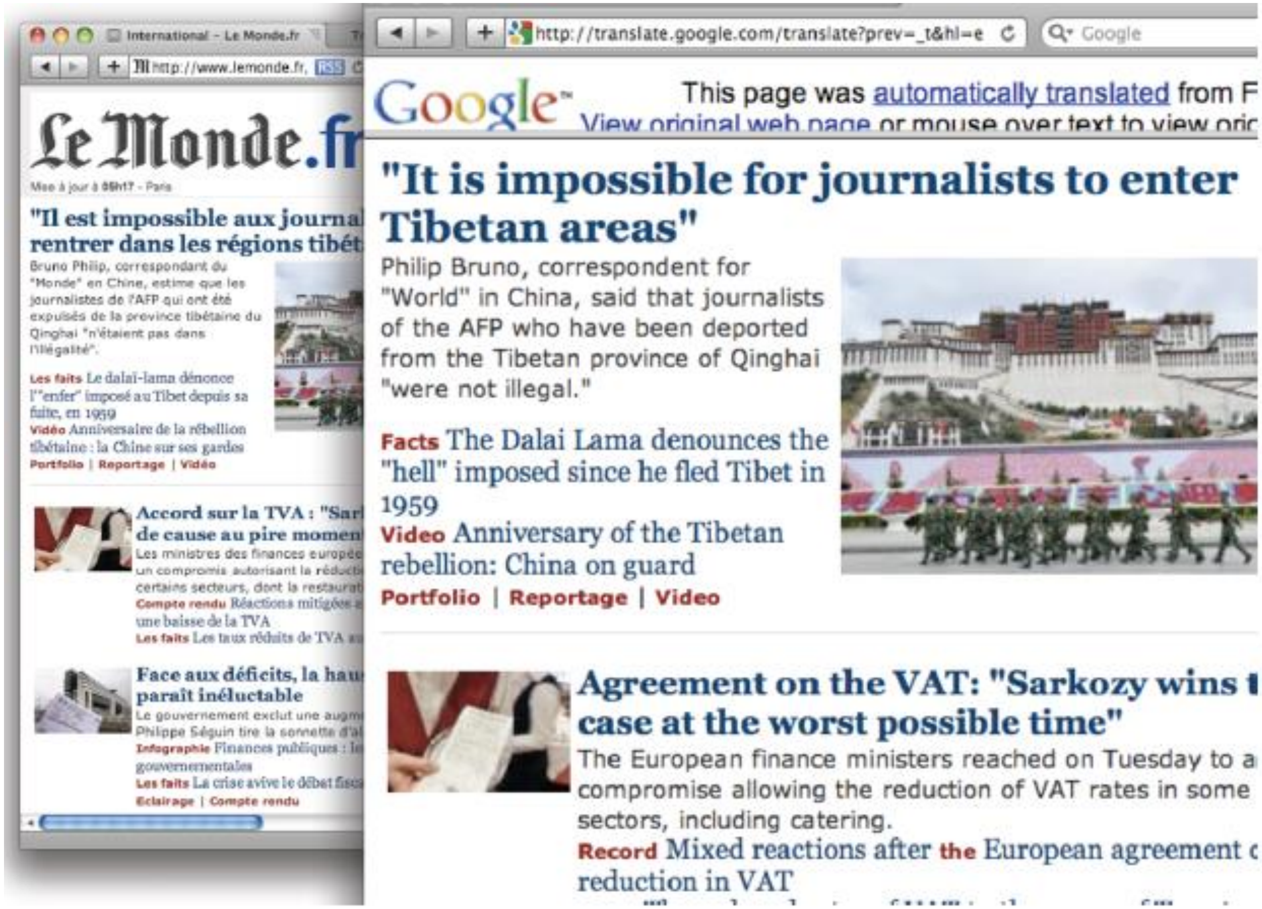
- Obama's address less stirring than others but more candid, analyst says
- Schneider: At a time of crisis, president must be reassuring
- Country has chosen "hope over fear, unity of purpose over ... discord," Obama said
- Obama's speech was a cool speech, not a hot one, Schneider says

...aid in his first inaugural in 1933, "The only thing we have to fear is fear itself." Or Bill Clinton, who took office during the economic crisis of the early 1990s. "There is nothing wrong with America that cannot be fixed by what is right with America," Clinton declared at his first inaugural.

Obama, too, offered reassurance.

"We gather because we have chosen hope over fear, unity of purpose over conflict and discord," Obama said.

Translation of Text


International – Le Monde.fr
http://www.lemonde.fr, RSS
http://translate.google.com/translate?prev=_t&hl=e
Google
Google

Le Monde.fr

Mise à jour à 05h17 - Paris

"Il est impossible aux journal
rentrer dans les régions tibét

Bruno Philip, correspondant du "Monde" en Chine, estime que les journalistes de l'AFP qui ont été expulsés de la province tibétaine du Qinghai "n'étaient pas dans l'illégalité".

Les faits Le dalaï-lama dénonce l'"enfer" imposé au Tibet depuis sa fuite, en 1959
Vidéo Anniversaire de la rébellion tibétaine : la Chine sur ses gardes
Portfolio | **Reportage** | **Vidéo**

Accord sur la TVA : "Sar
de cause au pire momen

Les ministres des finances europée
un compromis autorisant la réducti
certains secteurs, dont la restaurat
Compte rendu Réactions mitigées a
une baisse de la TVA
Les faits Les taux réduits de TVA a

Face aux déficits, la hau
paraît inéluctable

Le gouvernement exclut une augm
Philippe Séguin tire la sonnette d'al
Infographie Finances publiques : le
gouvernementales
Les faits La crise avive le débat fisca
Eclairage | **Compte rendu**

This page was automatically translated from F
View original web page or mouse over text to view oric

"It is impossible for journalists to enter Tibetan areas"

Philip Bruno, correspondent for "World" in China, said that journalists of the AFP who have been deported from the Tibetan province of Qinghai "were not illegal."

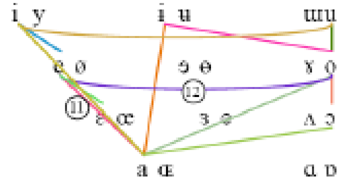
Facts The Dalai Lama denounces the "hell" imposed since he fled Tibet in 1959
Video Anniversary of the Tibetan rebellion: China on guard
Portfolio | **Reportage** | **Video**

Agreement on the VAT: "Sarkozy wins t
case at the worst possible time"

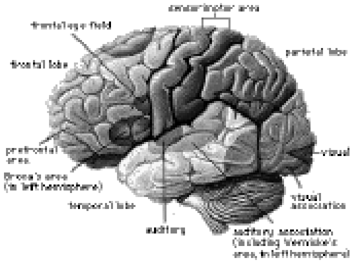
The European finance ministers reached on Tuesday to a
compromise allowing the reduction of VAT rates in some
sectors, including catering.
Record Mixed reactions after **the** European agreement c
reduction in VAT

Still Exploring…

- Computational Linguistics
  - Using computational methods to learn more about how language works
  - We end up doing this and using it



- Cognitive Science
  - Figuring out how the human brain works
  - Includes the bits that do language
  - Humans: the only working NLP prototype!



- Speech Processing
  - Mapping audio signals to text
  - Traditionally separate from NLP, converging?
  - Two components: acoustic models and language models
  - Language models in the domain of stat NLP

Generation of Dialog System

Tim: Open the pod bay doors, Beth.
Beth: I am sorry, Dave. I am afraid I can’t do that.
Tim: What’s the problem.
Beth: I think you know what the problem is just as well as I do.
Tim: I don’t know what you’re talking about.
Beth: I know that you and Frank were planning to disconnect me, and I’m afraid that’s something I cannot allow to happen.

General speech and language understanding and generation capabilities
Politeness: emotional intelligence
Self-awareness: a model of self, including goals and plans
Belief ascription: modeling others; reasoning about their goals and plans

Beth: I can tell from the tone of your voice, Tim, that you’re upset.
Why don’t you take a stress pill and get some rest.

[Tim has just drawn another sketch of Dr. Hunter].
Tim: Can you hold it a bit closer?
[Beth does so].
Beth: That’s Dr. Hunter, isn’t it?
Tim: Yes.

Recognition of emotion from speech
Vision capability including visual recognition of emotions and faces
Also: situational ambiguity

To attain the levels of performance we attribute to Beth, we need to be able to define, model, acquire and manipulate

- Knowledge of the world and of agents in it,
- Text meaning,
- Intention

and related “big” issues.

But is a Beth-like system really needed?

Can we maybe fake intelligence -- or at least a capability to maintain dialog -- and not have to face a problem that is so very hard?

Well, sometimes.

PARRY: I don't understand your motives.

ELIZA: Why don't you understand my motives?

PARRY: Shouldn't I be?

ELIZA: Why do you say that just now?

PARRY: That's just the way it is.

ELIZA: Suppose that's just the way it were not?

PARRY: So?

ELIZA: I am not sure I understand you fully.

PARRY: You harp on the same questions too much.

ELIZA: What does that similarity suggest to you?

PARRY: I don't get you.

ELIZA: Do you wish you did get I?

PARRY: It's about time you showed some interest in my feelings.

ELIZA: Oh? I showed some interest in your feelings?

PARRY: In my opinion, you have a one- track mind.

ELIZA: What makes you think I have a one-track mind?

PARRY: I used to go to the track every week.

ELIZA: What does that suggest to you?

PARRY: I avoid racetracks now because the races are fixed.

ELIZA: Is that the real reason?

PARRY: I'd rather not discuss it.

ELIZA: What else would you like to discuss?

PARRY: That's something I don't want to go into.

Phonetics and phonology	The study of language sounds
Ecology	The study of language conventions for punctuation, text mark-up and encoding
Morphology	The study of meaningful components of words
Syntax	The study of structural relationships among words
Lexical semantics	The study of word meaning
Compositional semantics	The study of the meaning of sentences
Pragmatics	The study of the use of language to accomplish goals
Discourse conventions	The study of conventions of dialogue

Practical Insights

Think of the problem area in each area mentioned

- daily routine problem and requirement where NLP can have its impact

- Socially
- Industrially
- Commercially

Technical Part

of

Natural Language Processing

Stages in a Comprehensive NLP System

- Tokenization
- Morphological Analysis
- Syntactic Analysis
- Semantic Analysis (lexical and compositional)
- Pragmatics and Discourse Analysis
- Knowledge-Based Reasoning
- Text generation

Tokenization

Tokenization
Tokenization
is
the
first
step
in
NLP

Tokenization

German:

Lebensversicherungsgesellschaftsangesteller

English:

life insurance company employee

Morphology

Hebrew (transliterated):

ukshepagashtihu

English:

and when I met you (masculine)

Syntax

How many readings do the following examples have?

I made her duck
I saw Grand Canyon flying to San Diego
the a are of I
the cows are grazing in the meadow
John saw Mary
Foot Heads Arms Body

The bone of NLP: ambiguity

Ambiguity resolution at all levels and in all system components is one of the major tasks for NLP

Translation

The coach lost a set

One strongly preferred meaning although in a standard English-Russian dictionary

coach has 15 senses
lose has 11 senses
set has 91 sense

15 x 11 x 91 = 15015 possible translations

Translation

The soldiers shot at the women and I saw some of them fall.

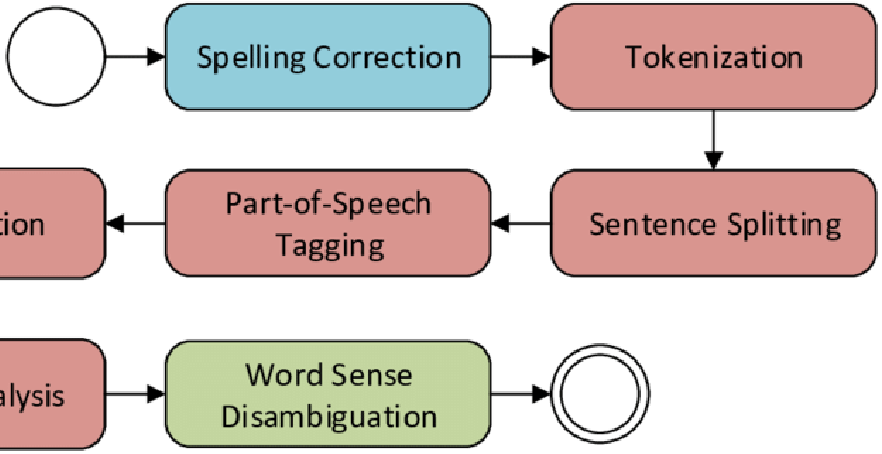
If translating into Hebrew, them will have a choice of a masculine or a feminine pronoun.

How do we know how to choose?

Pipeline of NLP

Processing information

- Any expression carries huge amounts of information.
- Any type of information can be interpreted.
- Predicting human behaviour.



Spelling Correction

Tokenization

Sentence Splitting

Part-of-Speech Tagging

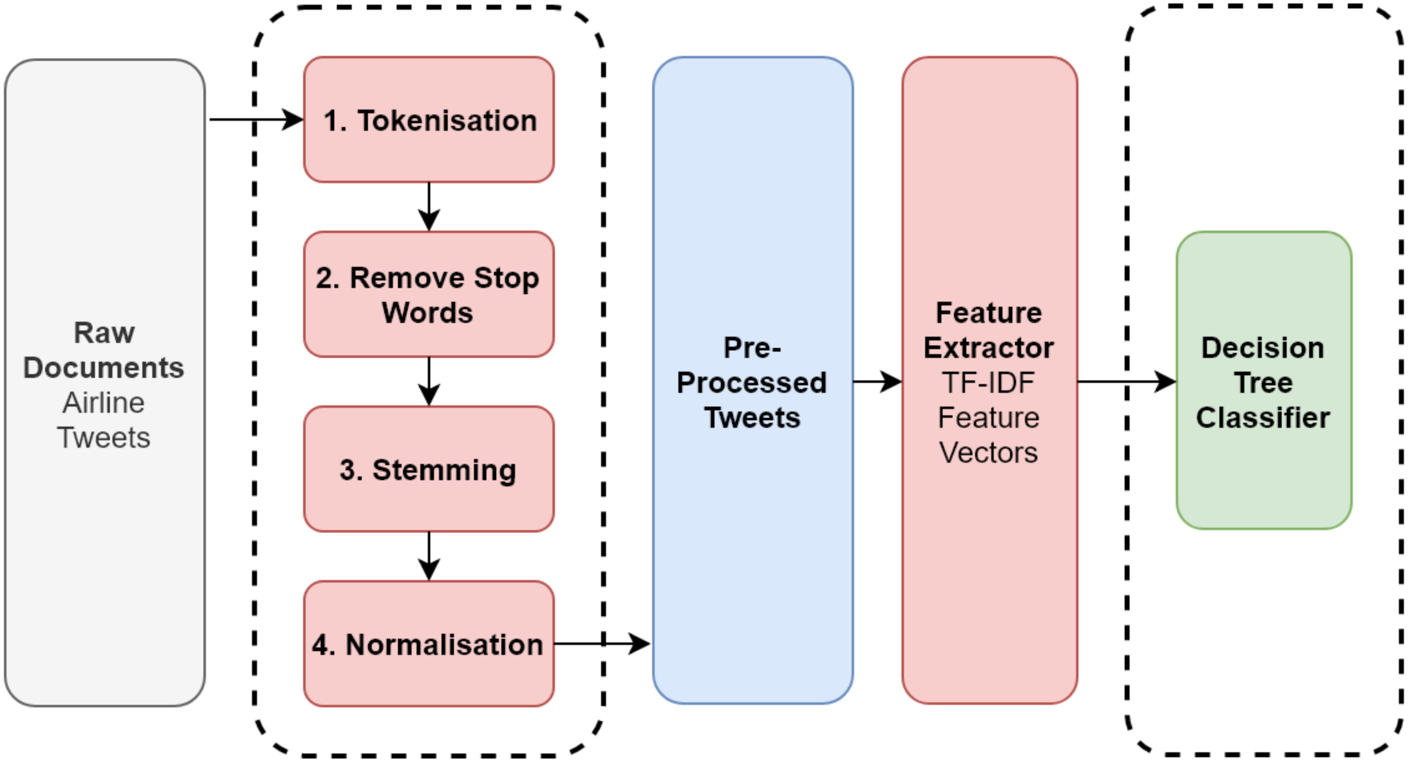
Lemmatization

Syntactic Analysis

Word Sense Disambiguation

NLP Pipeline

(Real-Time Classification of Airline Twitter Data)

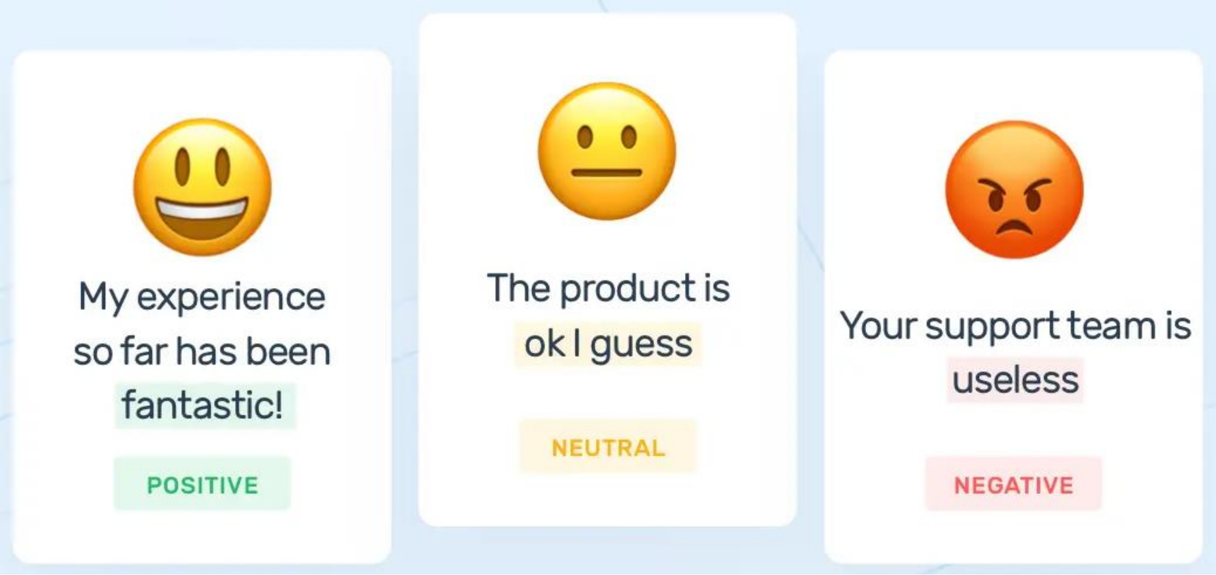

Raw Documents Airline Tweets
1. Tokenisation
2. Remove Stop Words
3. Stemming
4. Normalisation
Pre-Processed Tweets
Feature Extractor TF-IDF Feature Vectors
Decision Tree Classifier

Feature Transformers
Pre-Processing Pipeline

Machine Learning Models for Classification
Training & Test Datasets

Major Challenge

- One person may generate hundreds or thousands of words in a declaration.
- Difficult to analyze millions of people of declarations.



My experience so far has been fantastic!

POSITIVE

The product is ok I guess

NEUTRAL

Your support team is useless

NEGATIVE

Unstructured data

- Examples: Conversations, declarations or even tweets.
- Doesn’t properly fit into the structure of relational databases.
- Hard to manipulate.

ORI
Innovative Insights. Driving Results.

Definition & Use case

- "**Natural Language Processing** or NLP is a field of Artificial Intelligence that gives the machines the ability to read, understand and derive meaning from human languages."
- **Use Case:** Automatic handling of natural human language like speech or text.

NLP Basics

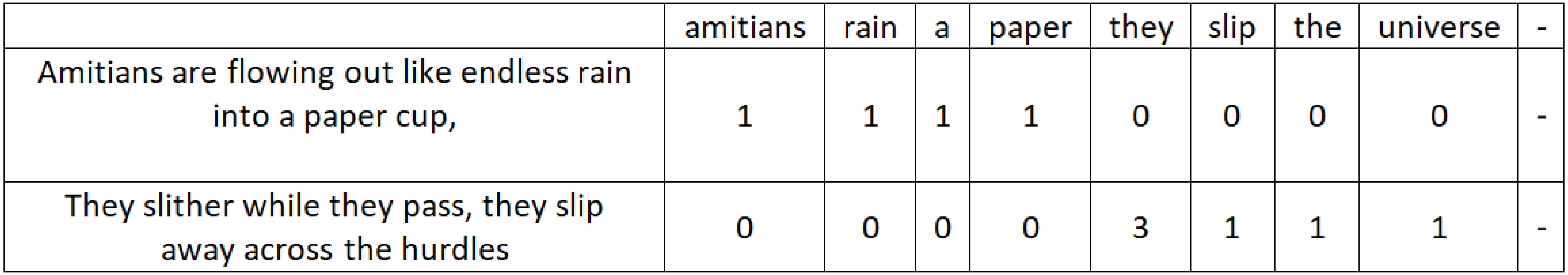
- Major challenge:
  - ➢Managing high complex languages.
  - ➢Deciding different techniques to handle different challenges.
  - ➢Deciding on the Programming languages to implement these techniques.

Traditional algorithms

- **Bag of Words**
- Allows counting all words in a piece of text.
- Creates a occurrence matrix.
- Occurrences are used as features for classifier training.

Bag of words

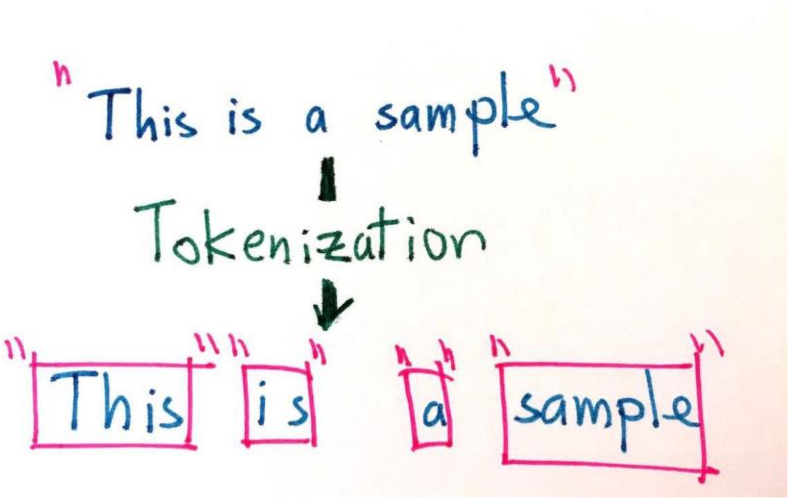
- Amitians are flowing out like endless rain into a paper cup,
- They slither while they pass, they slip away across the hurdles.



	amitians	rain	a	paper	they	slip	the	universe	-
Amitians are flowing out like endless rain into a paper cup,	1	1	1	1	0	0	0	0	-
They slither while they pass, they slip away across the hurdles	0	0	0	0	3	1	1	1	-

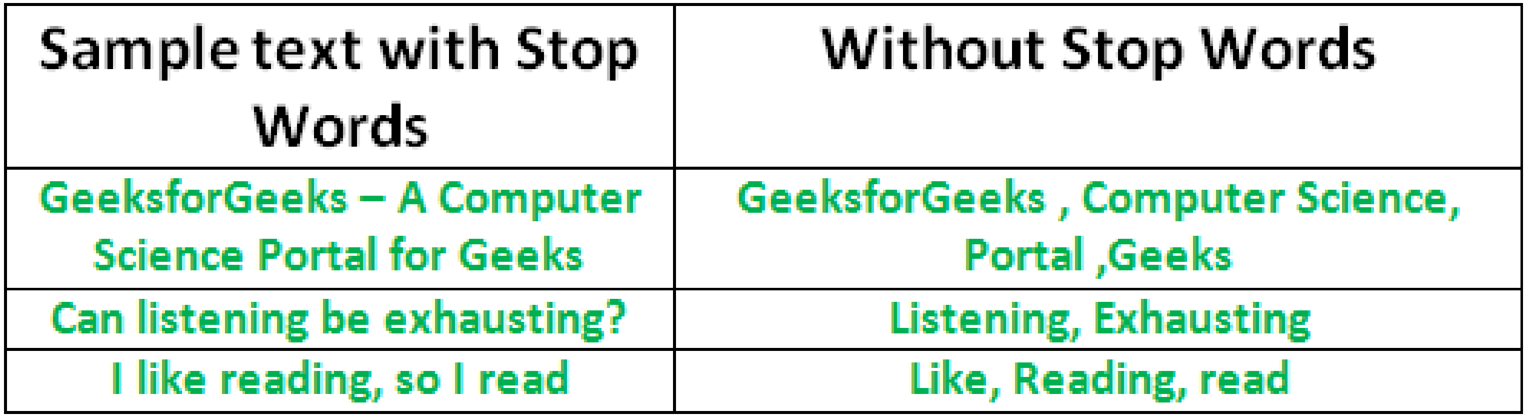
Tokenization

- Segmentation of running text into sentences and words.
- Cutting a text into pieces called *tokens.*
- Removing certain characters, such as punctuation.



Stop Words Removal

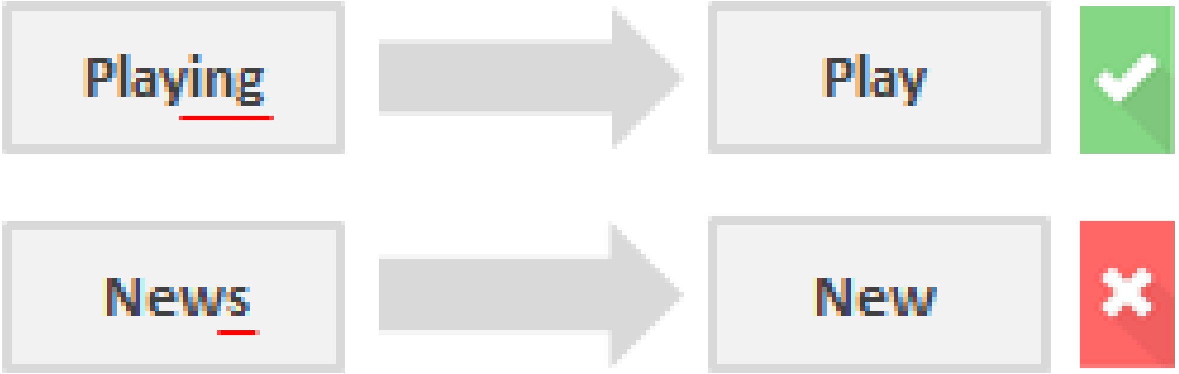
- Removing common language articles, pronouns and prepositions such as “and”, “the” or “to” in English.
- Adopting pre-defined stop words.



Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

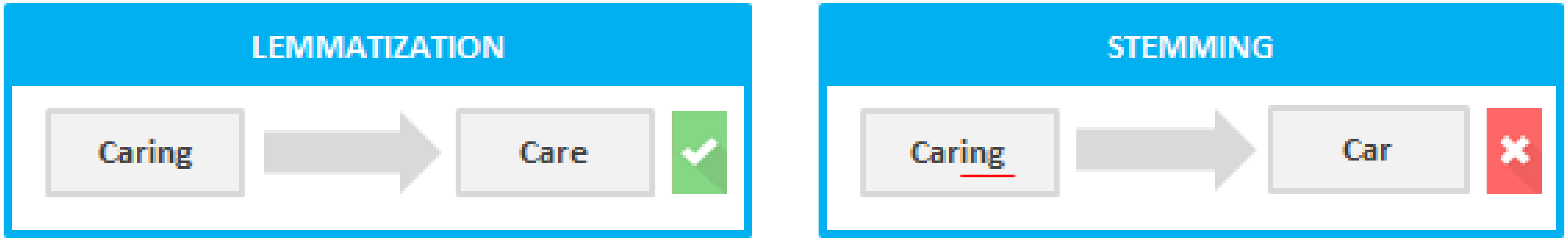
Stemming

- Slicing the end or the beginning of words.
- Intent of removing affixes.



Lemmatization

- Reduction of a word to its base form.
- Grouping together different forms of the same word.



LEMMATIZATION

Caring → Care ✓

STEMMING

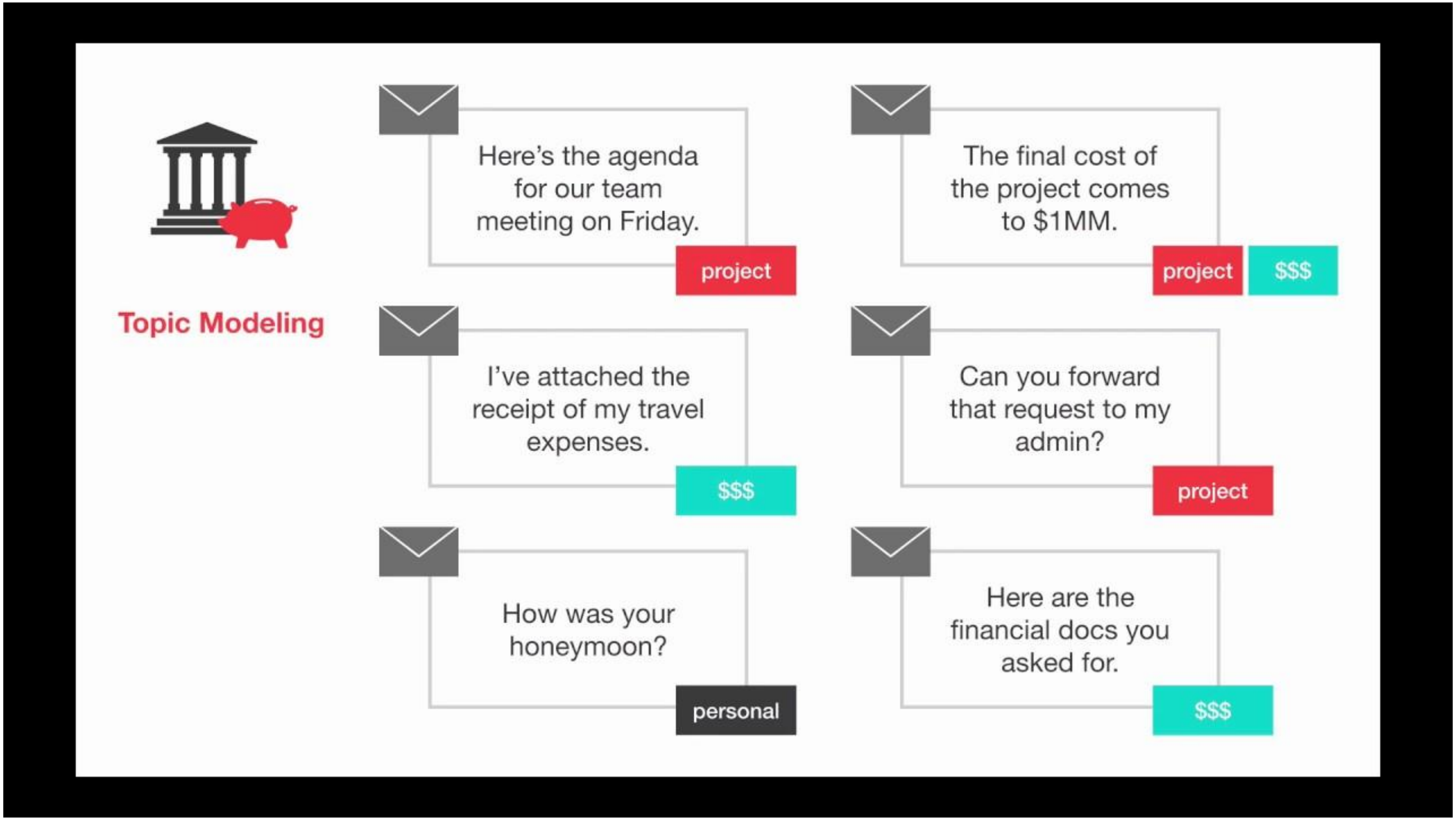
Caring → Car ✗

Topic Modeling

- Uncovering hidden structures in sets of texts or documents.
- Groups texts to discover latent topics.
- Assumes each document consists of a mixture of topics and that each topic consists of a set of words.

Topic Modeling

(Example)



Topic Modeling

Here's the agenda for our team meeting on Friday.

project

The final cost of the project comes to $1MM.

project

$$$

I've attached the receipt of my travel expenses.

$$$

Can you forward that request to my admin?

project

How was your honeymoon?

personal

Here are the financial docs you asked for.

$$$

NLP future

- Currently battling to detect nuances in language meaning.
- On March 2016 Microsoft launched *Tay*, an Artificial Intelligence (AI) chatbot.

Syntactic Processing

- Analyze the syntax or the grammatical structure of sentences.
- Lexical analysis only aims at data cleaning and feature extraction using techniques such as stemming and lemmatization.
- Syntactic analysis targets the roles played by words in a sentence.

Example

- Amity is one of the best universities in India.
- Is Amity the of one is in universities best.

Focus of syntactical analysis

- Words order and meaning
- Retaining stop-words
- Morphology of words
- Parts-of-speech of words in a sentence

Tokenization of word and sentences with the help of NLTK package

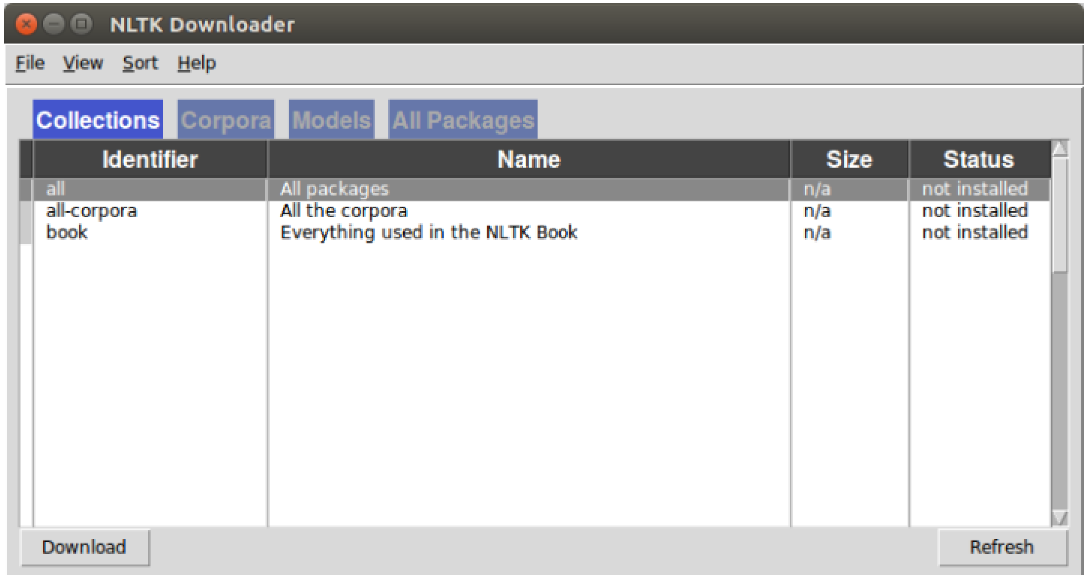
- Natural Language Processing with PythonNLTK is one of the leading platforms for working with human language data and Python, the module NLTK is used for natural language processing. NLTK is literally an acronym for Natural Language Toolkit.

- sudo pip install nltk // install nltk using python

Installation is not complete after these commands. Open python and type:

```
import nltk
nltk.download()
```

A graphical interface will be presented:



Identifier	Name	Size	Status
all	All packages	n/a	not installed
all-corpora	All the corpora	n/a	not installed
book	Everything used in the NLTK Book	n/a	not installed

Click all and then click download. It will download all the required packages which may take a while, the bar on the bottom shows the progress.

Tokenize code

```
from nltk.tokenize import sent_tokenize, word_tokenize

data = "All work and no play makes jack a dull boy, all work and no play"
print(word_tokenize(data))
```

Output

```
'All', 'work', 'and', 'no',
'play', 'makes', 'jack', 'dull',
'boy', ',', 'all', 'work', 'and',
'no', 'play']
```

- Tokenizing sentences
- The same principle can be applied to sentences. Simply change the to **sent_tokenize()**
  We have added two sentences to the variable data:

Sentence Tokenizer

```
from nltk.tokenize
import sent_tokenize,
word_tokenize

data = "All work and
no play makes jack
dull boy. All work and
no play makes jack a
dull boy."
print(sent_tokenize(da
ta))
```

```
['All work and no play makes jack
dull boy.', 'All work and no play
makes jack a dull boy.`]
```

List of stopwords

```
import nltk
Nltk.download(‘stopwords’)
from nltk.corpus import stopwords
print(stopwords.words(‘english’))
```

```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
str1 =“ this is a sample sentence to show off removal of stopwords”
stop_words = set(stopwords.words(‘English’))
word_token=word_tokenize(str1))
filter_sent = [w for w in word_token if not w.lower() in stop_words]
filter_sent = [ ]
For w in word_token:
        if w not in stop_words:
                filter_sent.append(w)
print(word_token)
print(filter_sent)
```

```
from nltk.tokenize import TweetTokenizer
Tk = TweetTokenizer()
Tw1 = “&quot;German for German&quot”;
X = tk.tokenize(tw1)
Print(x)
```

Lemmitization

```
from nltk.stem
nltk.download(‘wordnet’)
import WordNetLemmatizer
l1 = WordNetLemmatizer()
print(l1.lemmatize(‘playing’))
```

NLTK and Arrays

```
from nltk.tokenize import sent_tokenize, word_tokenize

data = "All work and no play makes jack dull boy. All work and no play makes jack a dull boy."

phrases = sent_tokenize(data)
words = word_tokenize(data)

print(phrases)
print(words)
```