Practical Applications of NLP

Universitat Pompeu Fabra Barcelona

Before we start...

Juan Soler Company

juan.soler@upf.edu

JUAN no JOAN.

55.408

NLP part

3 Theory classes

- Introduction and Applications
- Classic NLP
- Modern NLP

2 Labs (puntuables)

- Pos Tagging
- Author Profiling

1 Seminar (no puntuable)

Index

Overview

What?

How?

NLP Example Studies

Introduction

Natural Language Processing or NLP (not to be confused with Neuro Linguistic Programming) is a research field between AI and linguistics.

This field is concerned with the interactions between computers and humans, using the language of the latest.

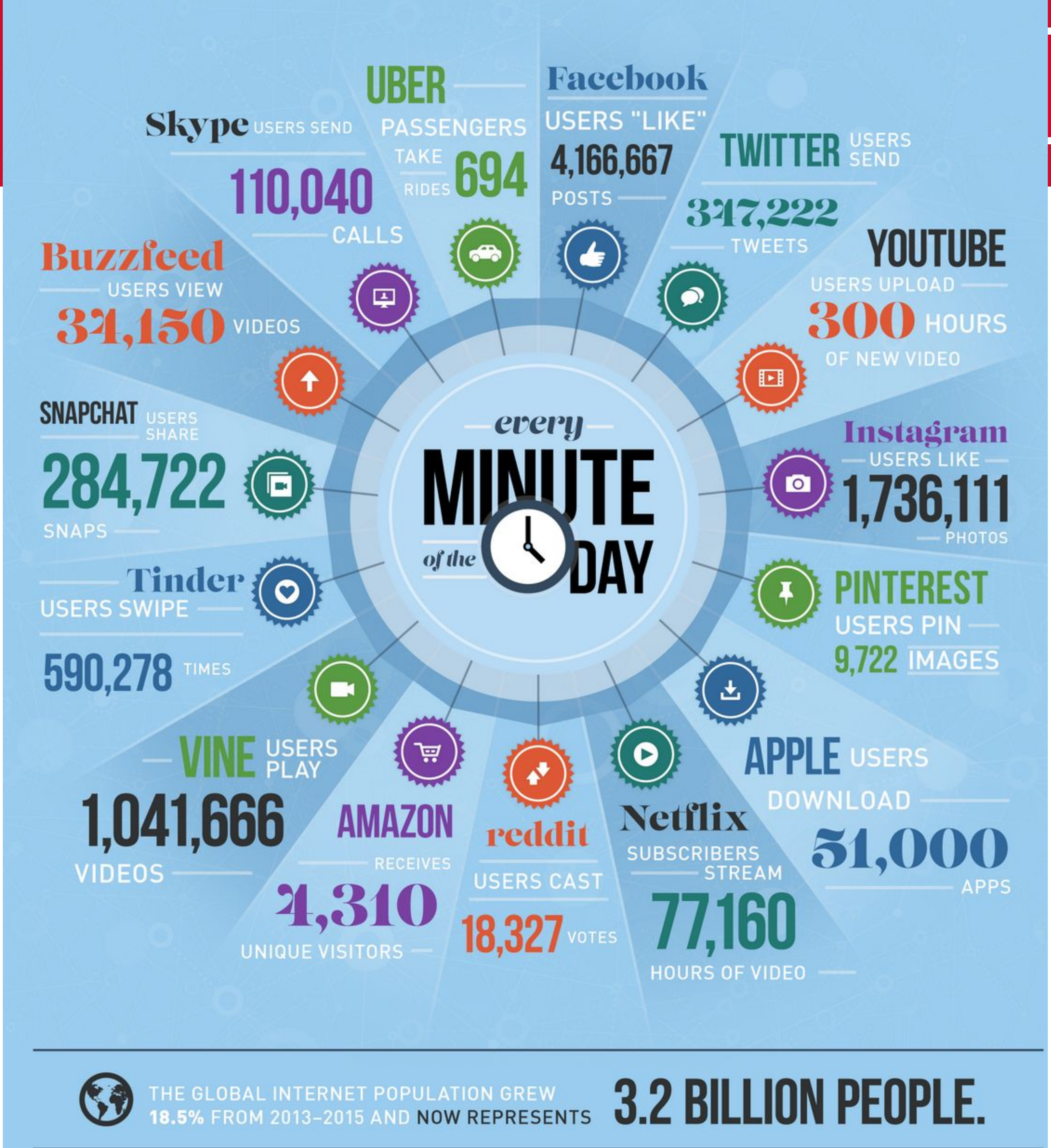
Introduction

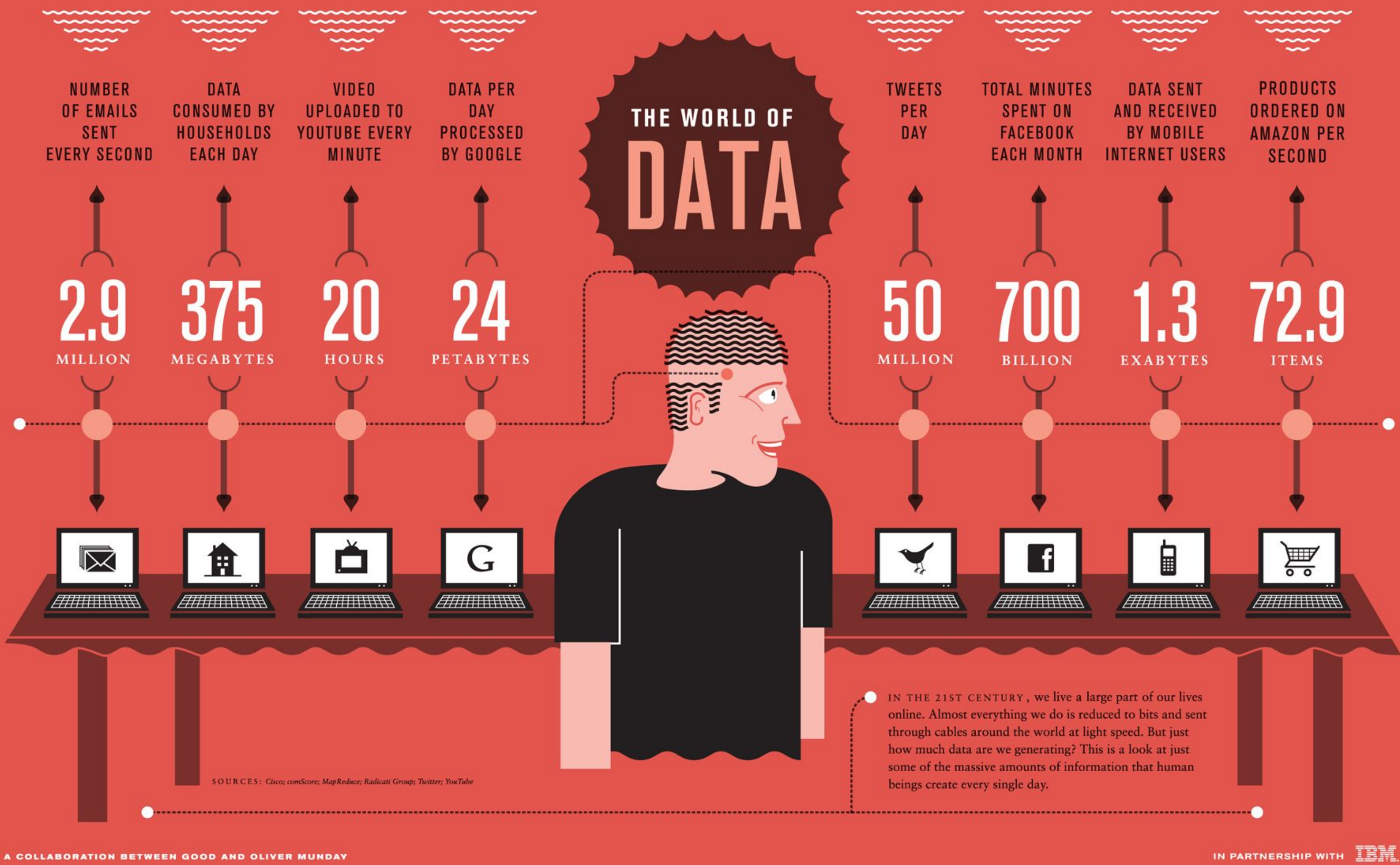
Why do we care?

Huge amounts of data in the net.

To make sense of it, we need to use NLP techniques.

upf. taln


every MINUTE of the DAY
Skype USERS SEND 110,040 CALLS
UBER PASSENGERS TAKE 694 RIDES
Facebook USERS "LIKE" 4,166,667 POSTS
TWITTER USERS SEND 347,222 TWEETS
YOUTUBE USERS UPLOAD 300 HOURS OF NEW VIDEO
Buzzfeed USERS VIEW 34,150 VIDEOS
SNAPCHAT USERS SHARE 284,722 SNAPS
Instagram USERS LIKE 1,736,111 PHOTOS
Tinder USERS SWIPE 590,278 TIMES
PINTEREST USERS PIN 9,722 IMAGES
VINE USERS PLAY 1,041,666 VIDEOS
AMAZON RECEIVES 4,310 UNIQUE VISITORS
reddit USERS CAST 18,327 VOTES
Netflix SUBSCRIBERS STREAM 77,160 HOURS OF VIDEO
APPLE USERS DOWNLOAD 51,000 APPS
THE GLOBAL INTERNET POPULATION GREW 18.5% FROM 2013–2015 AND NOW REPRESENTS 3.2 BILLION PEOPLE.


THE WORLD OF DATA
NUMBER OF EMAILS SENT EVERY SECOND
2.9 MILLION
DATA CONSUMED BY HOUSEHOLDS EACH DAY
375 MEGABYTES
VIDEO UPLOADED TO YOUTUBE EVERY MINUTE
20 HOURS
DATA PER DAY PROCESSED BY GOOGLE
24 PETABYTES
TWEETS PER DAY
50 MILLION
TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH
700 BILLION
DATA SENT AND RECEIVED BY MOBILE INTERNET USERS
1.3 EXABYTES
PRODUCTS ORDERED ON AMAZON PER SECOND
72.9 ITEMS
G
IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.
SOURCES: Cisco; comScore; MapReduce; Radicati Group; Twitter; YouTube
A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY
IN PARTNERSHIP WITH IBM

Who is using NLP?

Google

IBM

amazon

Microsoft

facebook

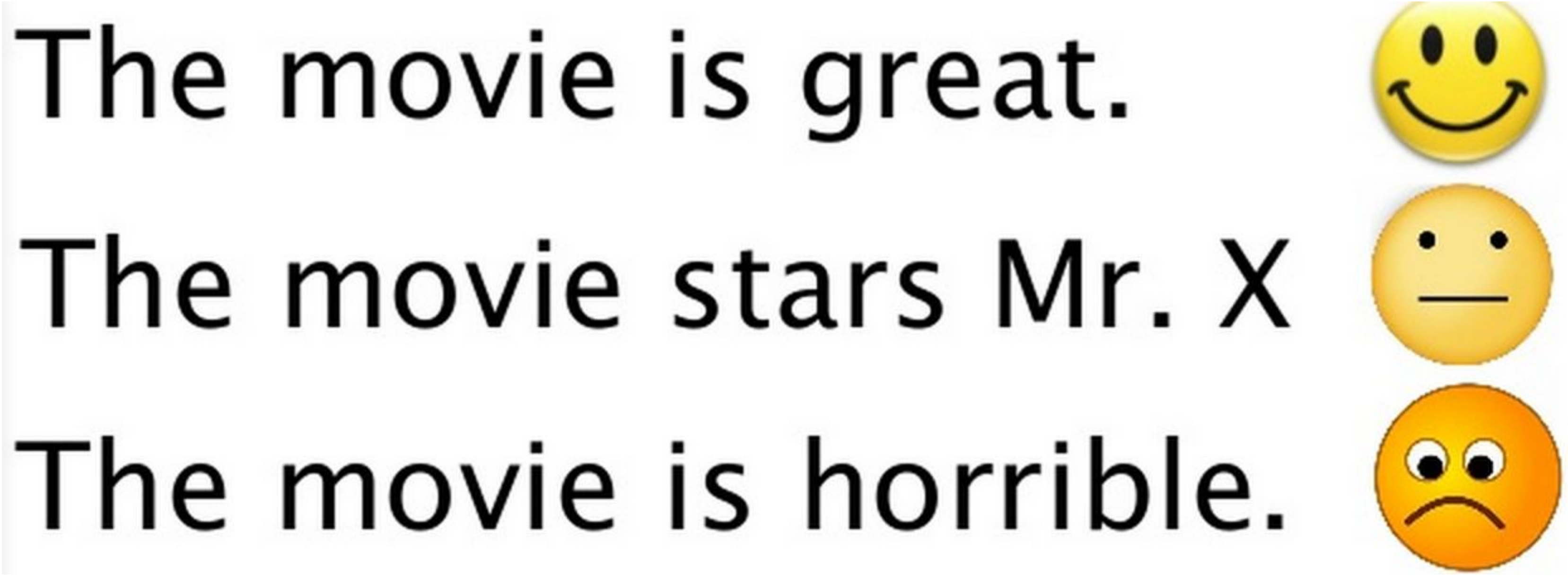
What?

Research Fields

- Automatic summarization
- Coreference resolution
- Machine translation
- Morphological segmentation
- Named entity recognition (NER)
- Natural language generation
- Parsing
- Question answering
- Sentiment analysis
- Speech recognition/segmentation
- Text Classification
- Word sense disambiguation
- …

Sentiment Analysis

Given a text input, determine if this text is «positive» or «negative».



Challenges:

Ambiguity, Irony/Sarcasm and Slang

“I absolutely love my phone company. What I like the most is their hability to steal my money without delivering”

“This game is looking absolutely sick”.

“TMB funciona siempre de manera sublime”.

Practical Applications?

Simple Example:

New and revolutionary TV by Sony.

They want to know what does their audience think.

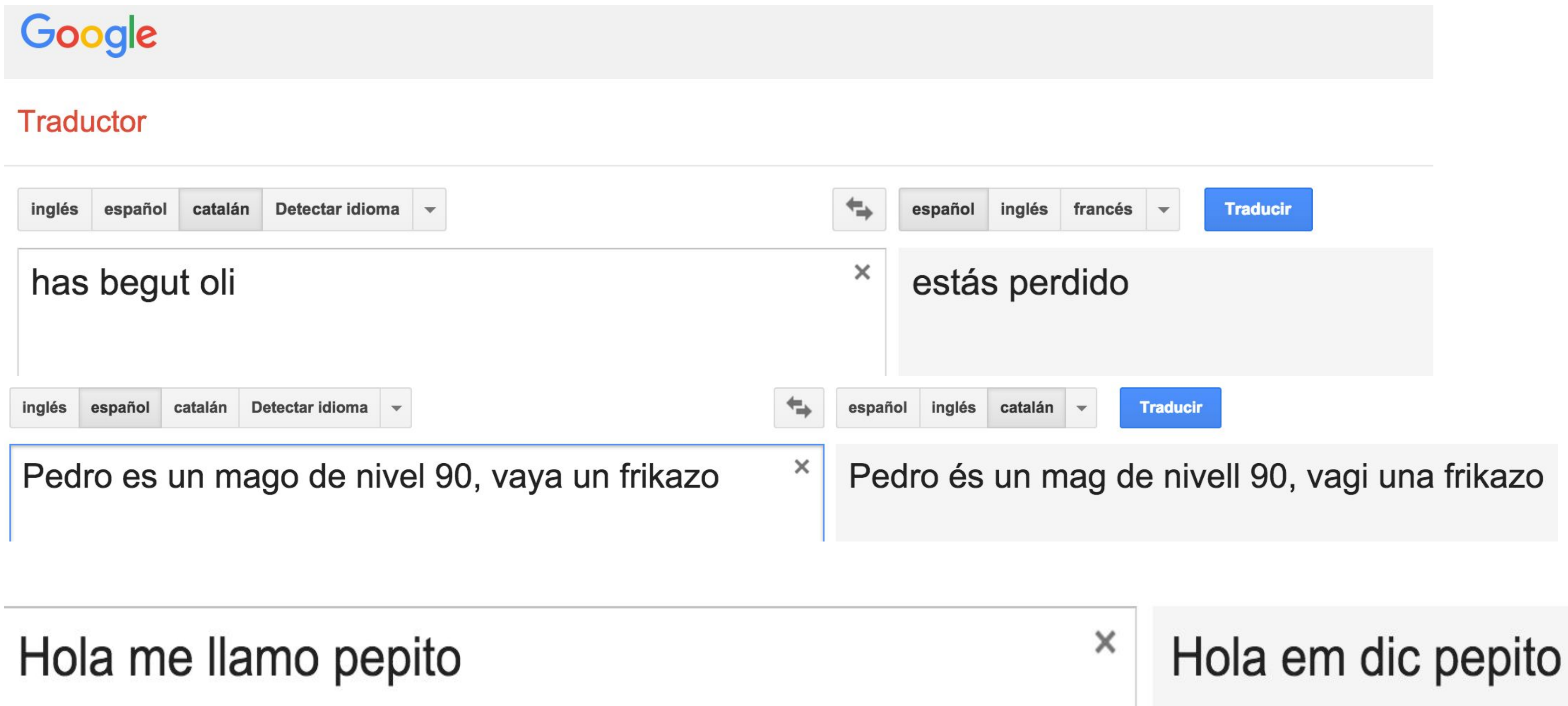
We can crawl amazon, and do sentiment analysis over the comments of people that have bought that TV set.

We can crawl the subreddit specialized in the matter.

We can also crawl specialized forums.

Machine Translation

Automatically translate text from one language to another.



Machine Translation

Old problem. This field started in the 1950s.

In 1955 this phrase was actually said: «in 5 to 10 years, this will be a solved problem».

…

They lied.

Text Classification

Given a text, classify it by:

- The topic that is being written about.
- The language in which it is written.
- Fiction vs non fiction.
- The author's gender (man or woman).
- If it is written by plumber or an engineer.
- …

Extra info:
http://www.scholarpedia.org/article/Text_categorization

Text Classification

Very important field in NLP and Information Retrieval.

Online Search Engines use these techniques: given some keywords, they return web pages (documents).

The keywords are processed and a classification process is performed (relevant documents vs non-relevant documents).

Text Classification

Other Applications:

Anti-Spam filter

Given an email, determine if it is a standard document or spam.

To do this, the mail must be analyzed and some features have to be extracted that can differentiate between spam or non-spam mails.

Dear former Participant of Interspeech ComParE,

It is with great pleasure that we announce the launch of the 10th event in this series:

The Interspeech 2018 Computational Paralinguistics Challenge
Atypical & Self-Assessed Affect, Crying & Heart Beats

Today, we launch the first two Sub-Challenges: Crying & Heart Beats.

Please find attached the Call for Participation, and the License Agreements for all Sub-Challenges – this year in one file and to be filled electronically.
If you would like to participate, you can fill all and send all as written in the PDF in one pass.
The homepage of the Challenge is to be found at: http://emotion-research.net/sigs/speech-sig/is18-compare

Thank you very much and hoping very much for your interest,

Hi juan.soler, My name is Anastasia and i'm writing you to tell you that you are super cute from your photos on Facebook. I myself am from Russia, but now I live in the USA. I want to get to know you more! If you have the same, email me, this is my email monikae3jabirgit@rambler.ru. Lets know each other better. Cheers, Anastasia

Hi juan.soler, the hottest man in the world! My name is Kseniya and i'm from Russia, but currently I live in the USA. I just wanted to let you know that I liked you from your photos and would like to know more about you. Let me know if you would like to get in touch, here is my email nita0n0ewing@o2.pl. Cheers, Kseniya

BANK OF AFRICA BENIN <"boa."@arrow.ocn.ne.jp> 5 feb. (hace 3 días)

para

¿Por qué está en Spam este mensaje? Porque se parece a otros mensajes detectados por nuestros filtros de spam. Más información

inglés > español Traducir mensaje Desactivar para: inglés

Dear Sir/Madam,

This is to officially inform you that we have verified your contract /inheritance file and found out that why you have not received your payment is because you have not fulfilled the obligations given to you in respect of your fund.

We have scheduled your transfer to be completed under 14 official banking days, that was because your payment file worth of $7,800,000.00 has less than 21 days to expire in this bank and when it expires, the fund will go into the Federal Government treasury account.

You will now receive your fund through online banking system and to receive your funds through online banking, you need to reply back with your information as urgent as possible to ensure you are the true beneficiary.
1. Full Name:.........
2. Address:.....
3. Country:.........
4. Phone number:.....

Yours sincerely,
Dr. Kendrick Lugan
Tel/ +229-6252-1741

Anti-Spam Filter

What kind of features could be useful?

Usage of currency symbols ($ €), exclamations, percentages.

Some keywords such as (money, penis, enlargement, nigerian-prince, discount...).

Url (of links in the mail or of the sender address) Analysis.

Number of links.

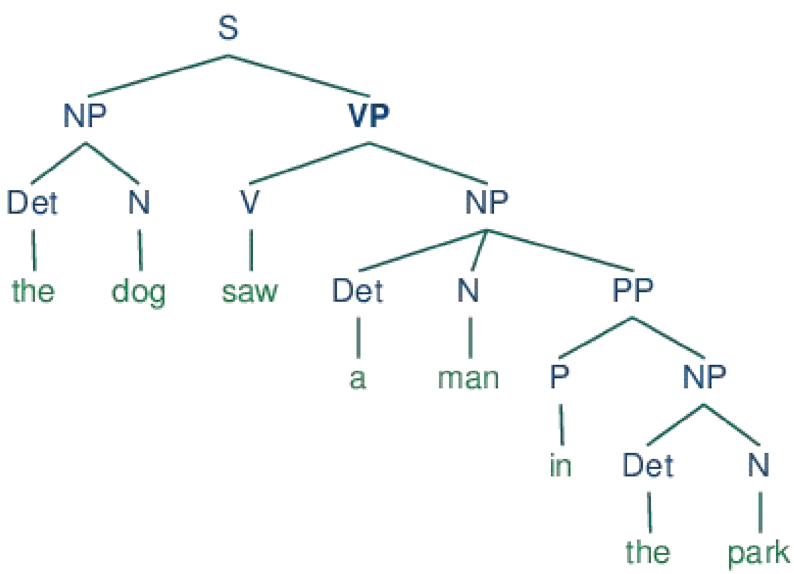
Upper cased characters

Parsing

Given a text, a parser analyzes it and generates a structure that can be of different kinds:

- Shallow
- Dependency Based
- Discursive
- …

Syntactic Parser



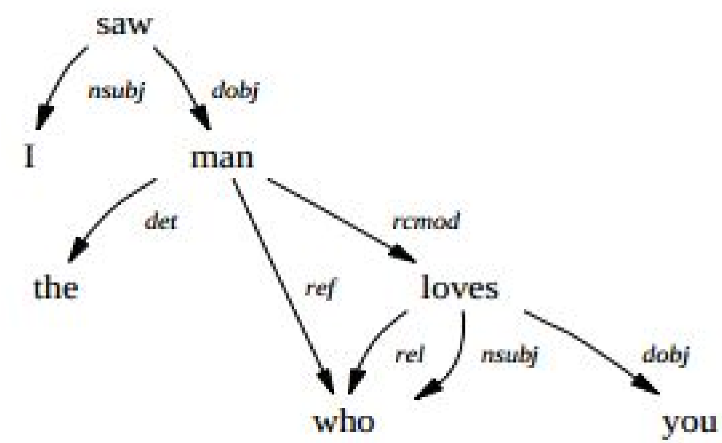
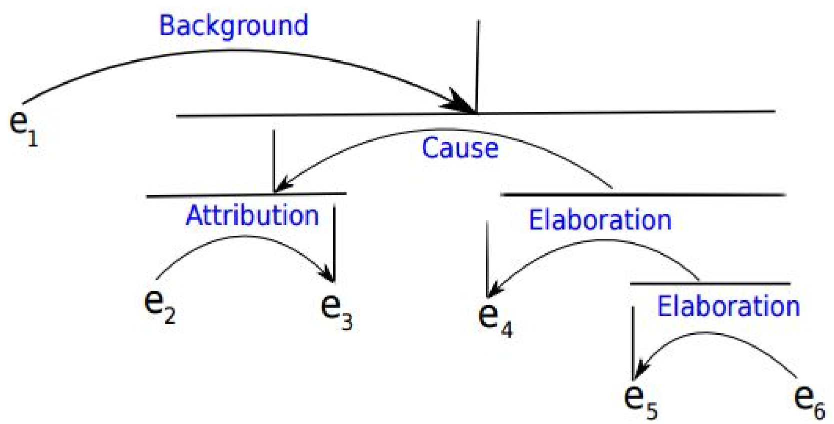


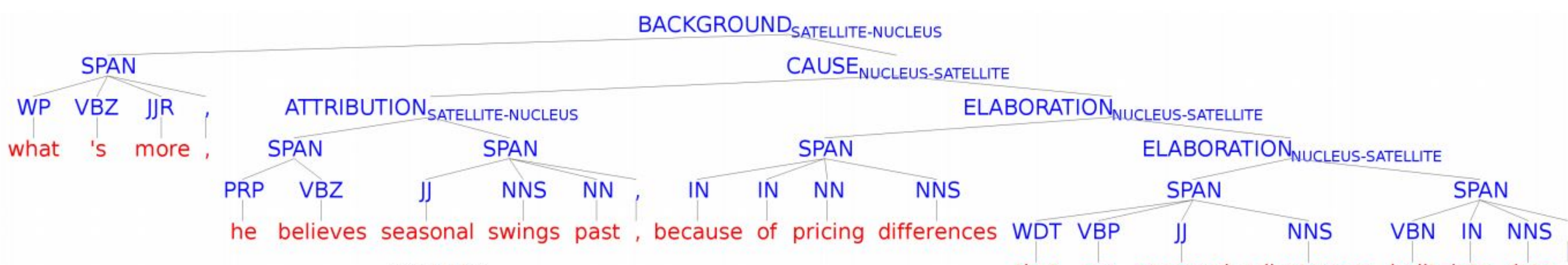
Figure 3: An example of a typed dependency parse for the sentence "I saw the man who loves you".

Shallow Parser Output

Dependency Parser Output

Discourse Parsing





(a) **JRN**

Summarization

Extractive

Given a text input, extract the most relevant phrases of it, the ones that summarize better the text.

Abstractive

Given a text, analyze it and generate a summary automatically, without using parts of the input text.

Coreference Resolution

Detect coreferences in a text

- Two or more terms are coreferent if the refer to the same concept.

Ex:

Peter went fishing. He was a big fan of it. Peter saw Mary. The boy said hi to her with a great smile on his face. Lola's daughter waved back without smiling. He was deep in the friendzone.

Named entity recognition

Given a text, identify and classify the elements that correspond to one of the next pre-defined categories:

- PERSON, LOCATION, ORGANIZATION, TIME, MONEY, PERCENTAGE.

Ex:

In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences. He settled in the U.S., becoming an American citizen in 1940. On the eve of World War II, he endorsed a letter to President Franklin D. Roosevelt alerting him to the potential development of "extremely powerful bombs of a new type" and recommending that the U.S. begin similar research. This eventually led to what would become the Manhattan Project. Einstein supported defending the Allied forces, but largely denounced using the new discovery of nuclear fission as a weapon. Later, with the British philosopher Bertrand Russell, Einstein signed the Russell-Einstein Manifesto, which highlighted the danger of nuclear weapons. Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

Tag colours:

LOCATION TIME PERSON ORGANIZATION MONEY PERCENT DATE

Morphological Segmentation/Analysis

- ***Morphology*** is the field of linguistics that studies the internal structure of words. (Wikipedia)
- A ***morpheme*** is the smallest linguistic unit that has semantic meaning (Wikipedia)
  - e.g. "carry", "pre", "ed", "ly", "s"
- Morphological analysis is the task of segmenting a word into its morphemes:
  - carried ⇒ carry + ed (past tense)
  - independently ⇒ in + (depend + ent) + ly
  - Googlers ⇒ (Google + er) + s (plural)
  - unlockable ⇒ un + (lock + able) ?

    ⇒ (un + lock) + able ?

Natural language generation

Field that is about generating natural language, using a knowledge base where information about concepts and relations within this concepts can be found, as well as the domain about which we want to write.

The system queries the knowledge base and translates the data that is represented in the knowledge base to natural language, trying to write like a human being.

Ex:

Siri

The input is voice. Siri analyzes the inputs and tries to know what did the user say. Then she responds according to the information she finds in their knowledge base about what the user said. It is also a dialog manager apart from a nl generator.

Question answering

Specialized systems that answer the questions a human being is asking about a specific domain.

The input of these systems can be voice or text-based.

They use a knowledge base that has specific information about the domain they are operating on. Using NL generation, they can answer in a human-like way.

A generalization of this kind of systems are **Dialog Managers**, that not only they answer questions but they can also maintain a normal conversation with a human being.

Cleverbot
Eliza

Cleverbot: I don't thinkI anyone wants to be my brothers girlfriend.

User: What's he look like?

Cleverbot: Who?

User: Your brother?

Cleverbot: I'm an only child.

User: You just said you had a brother.

Cleverbot: He died.

Cleverbot

Cleverbot, should I get a German Shepherd or a Labrador?

No get a tattoo of some pancakes instead.

Okay.

Cleverbot

cybersex

Ok.

you start

No, you start.

takes off shirt

Takes over russia.

User: Go! Pikachu!

Cleverbot: Pikachu!

User: Pikachu, use Thunderbolt!

Cleverbot: Pikachu, use quick attack!

User: That's enough, Pikachu! Come back! Go, Charizard!

Cleverbot: Go Pikachu!

User: No, Pikachu is my Pokemon.

Cleverbot: Pikachu will never walk again.

User: I....what?

Cleverbot: You heard me.

Speech Recognition/Segmentation

These research fields use voice as input. The goal is to recognize what is the human saying to transcribe his/her voice to written text.

One of the main challenges is to correctly split the different components of a phrase: determine where a word/syllable /phoneme starts and where it ends.

This can be hard due to different intonations, dialects, speech impediments, background noise… These are one of the most complex tasks in NLP.

Word Sense Disambiguation

Given a word, which of its senses is being used in a text?

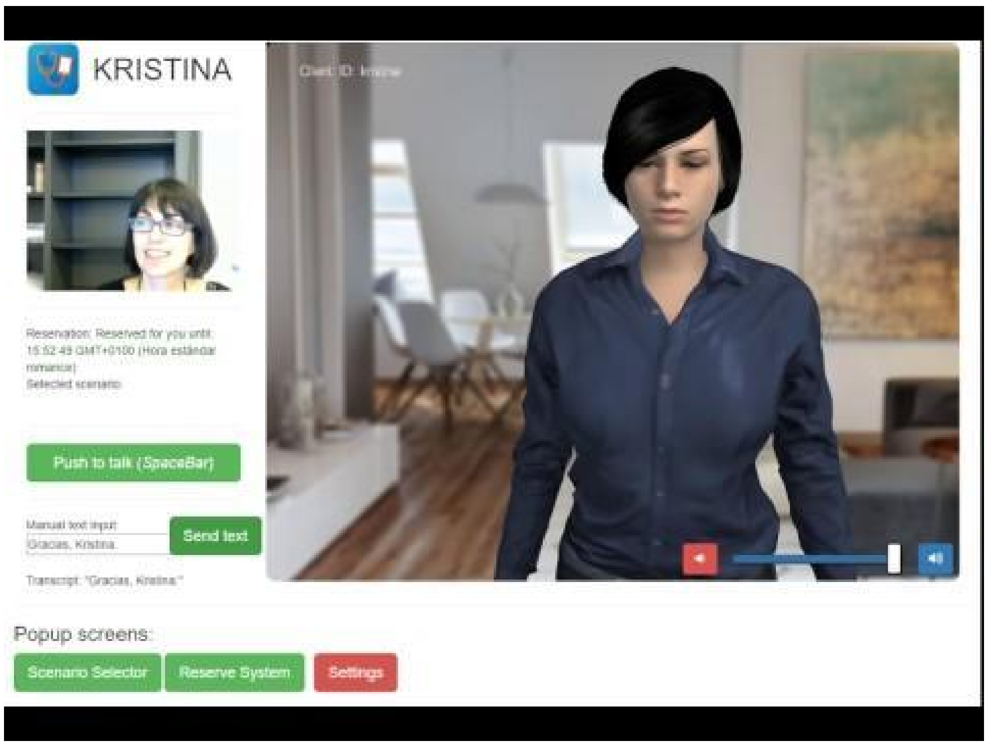
My dog loves to **run** about in the park

(move quickly with legs) ?
(manage) ?
(offer a service) ?
(period of time) ?
(race) ?

There are words with many senses and the context needs to be analyzed to discover which one is being used.

Kristina Project

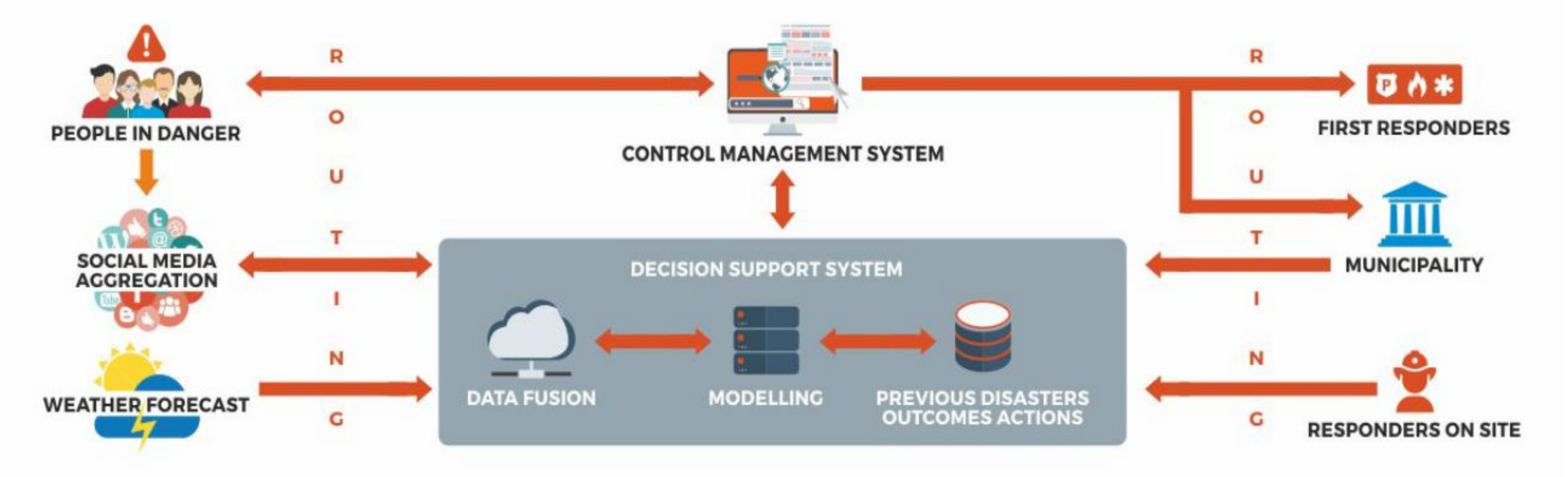
- European project in which lots of entities are involved (UPF being one of them).
- 3D avatar that helps immigrants communicate with doctors using voice.



KRISTINA's overall objective is to research and develop technologies for a human-like socially competent and communicative agent that is run on mobile communication devices and that serves for migrants with language and cultural barriers in the host country as a trusted information provision party and mediator in questions related to basic care and healthcare.

BeAware Project

beAWARE proposes an integrated solution to support forecasting, early warnings, transmission and routing of the emergency data, aggregated analysis of multimodal data and management the coordination between the first responders and the authorities. Our intention is to rely on platforms, theories and methodologies that are already used for disaster forecasting and management and add the elements that are necessary to make them working efficiently and in harm under the same objective.



How?

Rule-based systems

Machine Learning

Rule-based System

This approach can be useful in some cases.

Valid approach for problems in which all the possibilities that can happen are clear.

The question that needs to be answered in order to know if it can be solved with a rule based system is:

Can we write a list of rules that can model every possible situation?

Ex:

Information System:

If question 1 -> answer1

If question 2 -> answer2

...

Rule-based Systems

Ex2:

Tokenization.
Given a sentence, return the list of words that are part of it.

Sentence Splitting.
Given a text, return the list of sentences that compose it.

Using a finite list of rules, this problems can be solved very efficiently.

Rule-based Systems

Advantages:

No need for training data.

Can solve problems very efficiently.

Disadvantages:

If we are facing a big problem, writing thousand of rules to solve it, does not scale well (and it is not smart).

It has to be a specific kind of problem, sufficiently bounded to be solved this way.

Machine Learning

Statistic Approach

Branch of artificial intelligence that develops algorithms and techniques that help computers learn automatically.

It is about creating programs that are able to generalize knowledge by consuming unstructured data in the form of «solved» examples. It is a knowledge-induction process.

What people think machine learning is

What it really is

$$
\begin{aligned}
&\int_0^1 \mathrm{d}\xi \frac{\xi[(1-\xi)\mathbf{x}' + \xi\mathbf{x}''] \cdot \Delta\mathbf{x}}{|(1-\xi)\mathbf{x}' + \xi\mathbf{x}''|^3} \operatorname{erf}\left[\frac{|(1-\xi)\mathbf{x}' + \xi\mathbf{x}''|}{\sqrt{2\mathrm{i}\Lambda\xi(1-\xi)}}\right] \\
&\quad = \sqrt{\frac{2}{\pi\mathrm{i}\Lambda}} \int_0^1 \mathrm{d}\alpha \int_0^1 \mathrm{d}\xi \sqrt{\frac{\xi}{1-\xi}} \exp\left[-\frac{\alpha^2[(1-\xi)\mathbf{x}' + \xi\mathbf{x}'']^2}{2\mathrm{i}\Lambda\xi(1-\xi)}\right] \\
&\quad\quad \times \frac{[(1-\xi)\mathbf{x}' + \xi\mathbf{x}''] \cdot \Delta\mathbf{x}}{[(1-\xi)\mathbf{x}' + \xi\mathbf{x}'']^2} \\
&\quad = \sqrt{\frac{2\pi}{\mathrm{i}\Lambda}} \int_0^1 \mathrm{d}\alpha \left\{ \exp\left[\frac{\alpha^2\Delta\mathbf{x}^2}{2\mathrm{i}\Lambda}\right] \operatorname{erfc}\left[\frac{\alpha(r' + r'')}{\sqrt{2\mathrm{i}\Lambda}}\right] \right. \\
&\quad\quad \left. - \frac{1}{r''}\sqrt{\frac{r'r'' + \mathbf{x}' \cdot \mathbf{x}''}{2}} \operatorname{erfc}\left(\alpha\sqrt{\frac{r'r'' + \mathbf{x}' \cdot \mathbf{x}''}{\mathrm{i}\Lambda}}\right) \right\} \\
&\quad = \sqrt{\frac{2\pi}{\mathrm{i}\Lambda}} \int_0^1 \mathrm{d}\alpha \exp\left[\frac{\alpha^2\Delta\mathbf{x}^2}{2\mathrm{i}\Lambda}\right] \operatorname{erfc}\left[\frac{\alpha(r' + r'')}{\sqrt{2\mathrm{i}\Lambda}}\right] \\
&\quad\quad - \frac{1}{r''}\left[1 - \exp\left[-\frac{r'r'' + \mathbf{x}' \cdot \mathbf{x}''}{\mathrm{i}\Lambda}\right] + \sqrt{\frac{\pi(r'r'' + \mathbf{x}' \cdot \mathbf{x}'')}{\mathrm{i}\Lambda}} \operatorname{erfc}\left(\sqrt{\frac{r'r'' + \mathbf{x}' \cdot \mathbf{x}''}{\mathrm{i}\Lambda}}\right)\right],
\end{aligned}
\tag{32}
$$

Machine Learning Basics

Concepts:

Training and Testing Corpus.

Features.

Labels.

Machine Learning Algorithms.

Evaluation.

Training Set: set of instances labeled correctly. It can be seen as a set of «solved» examples.

Test Set: set of instances that will be used to test the extracted knowledge of our system and its predictions.

An Instance is the vectorial representation of one of the inputs as a **feature** multidimensional vector.

A **feature** is a characteristic that was extracted from the input and that it is considered to be relevant to characterize it.

A **label** is the category of the input, its correct solution. If we are trying to classify cats vs tigers, the label will indicate that an instance is a cat or a tiger.

[12, 33, 44, 7]
[f1,f2,f3,f4]
Label=0

Train.
While i < num instances train:

Test
While i < num inst test:

[-1, 3, 14, 72]
[f1,f2,f3,f4]
Label=?

ML Algorithm

Prediction.
Compare the prediction
with the real value.
Compute the accuracy.

ML Algorithms learn by extracting knowledge from «solved» examples.

Given an unlabeled instance, the algorithm uses the extracted knowledge from the training set (each ML algorithm does this in a different way) and makes a prediction.

Kinds of ML Problems

Supervised Machine Learning

We have a training corpus correctly labeled (a good number of «solved» examples).

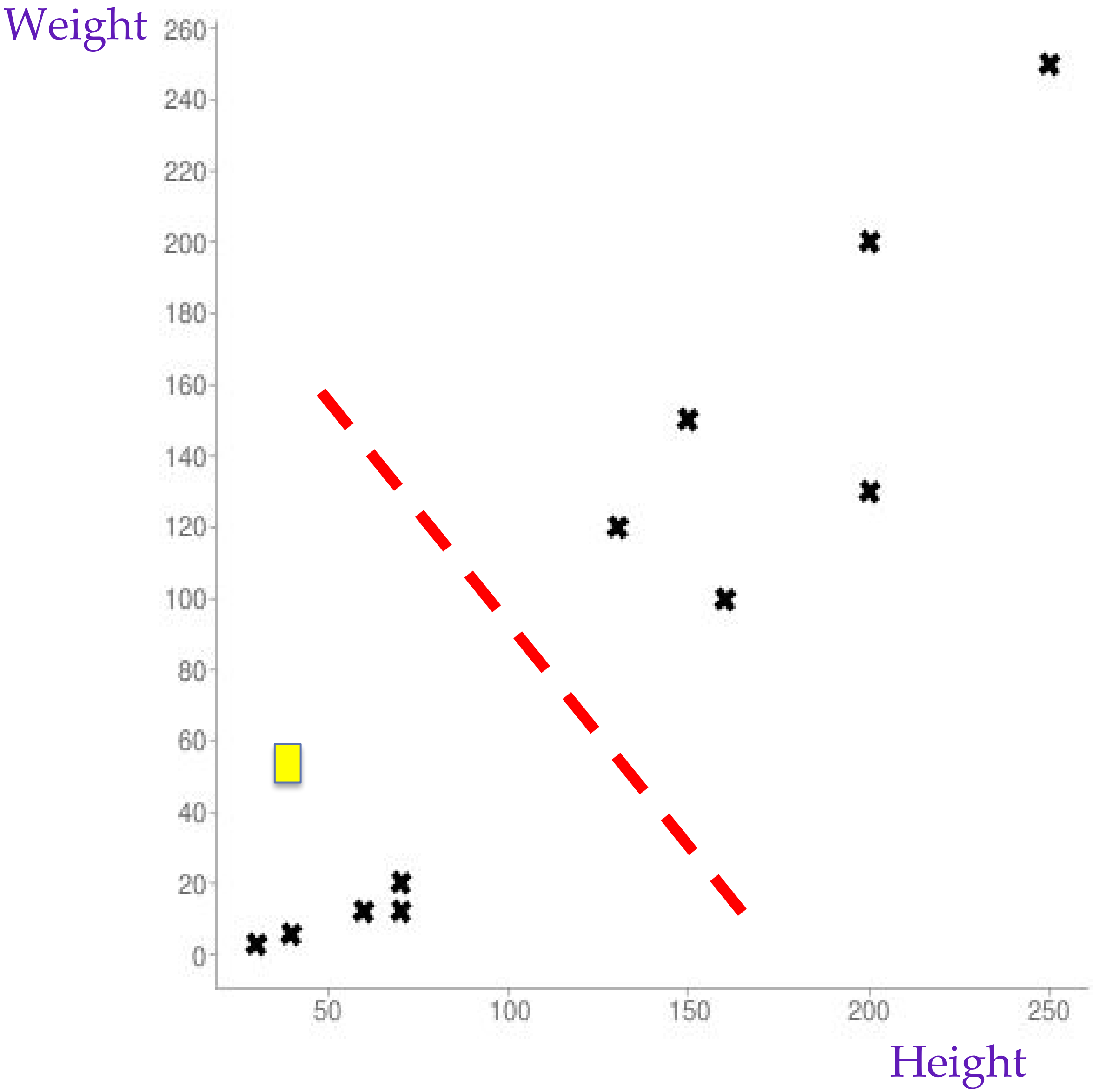
Semi Supervised Learning:

We have few training data and lots of unlabeled data. The idea is to use the unlabeled instances to help the algorithm learn in the training step.

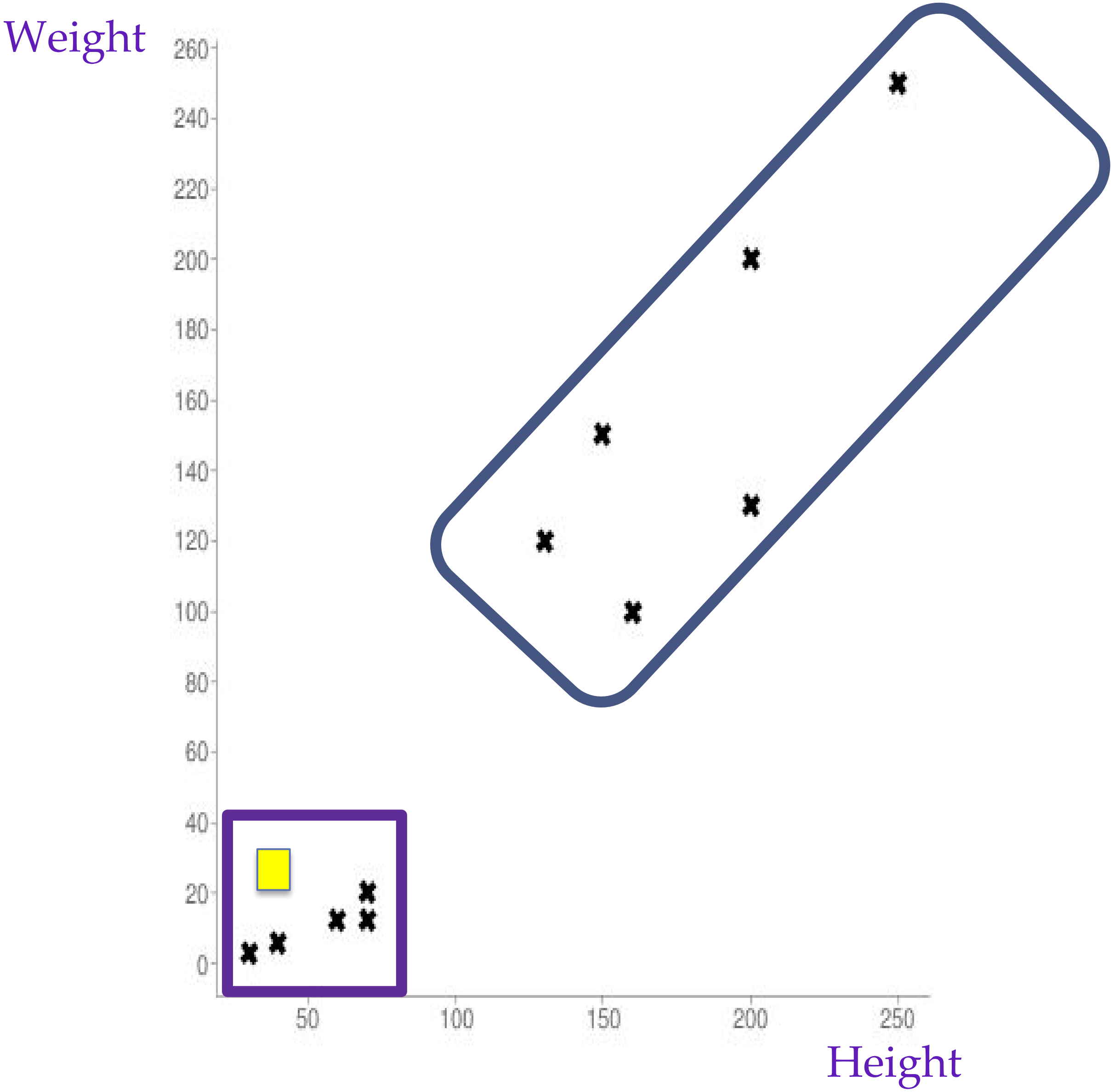
Unsupervised Learning (or clustering):

We don’t have labels. We will group the instances that are most similar. We can only say that these instances are similar to one another.

Cat or Tiger? (Supervised)



Unsupervised

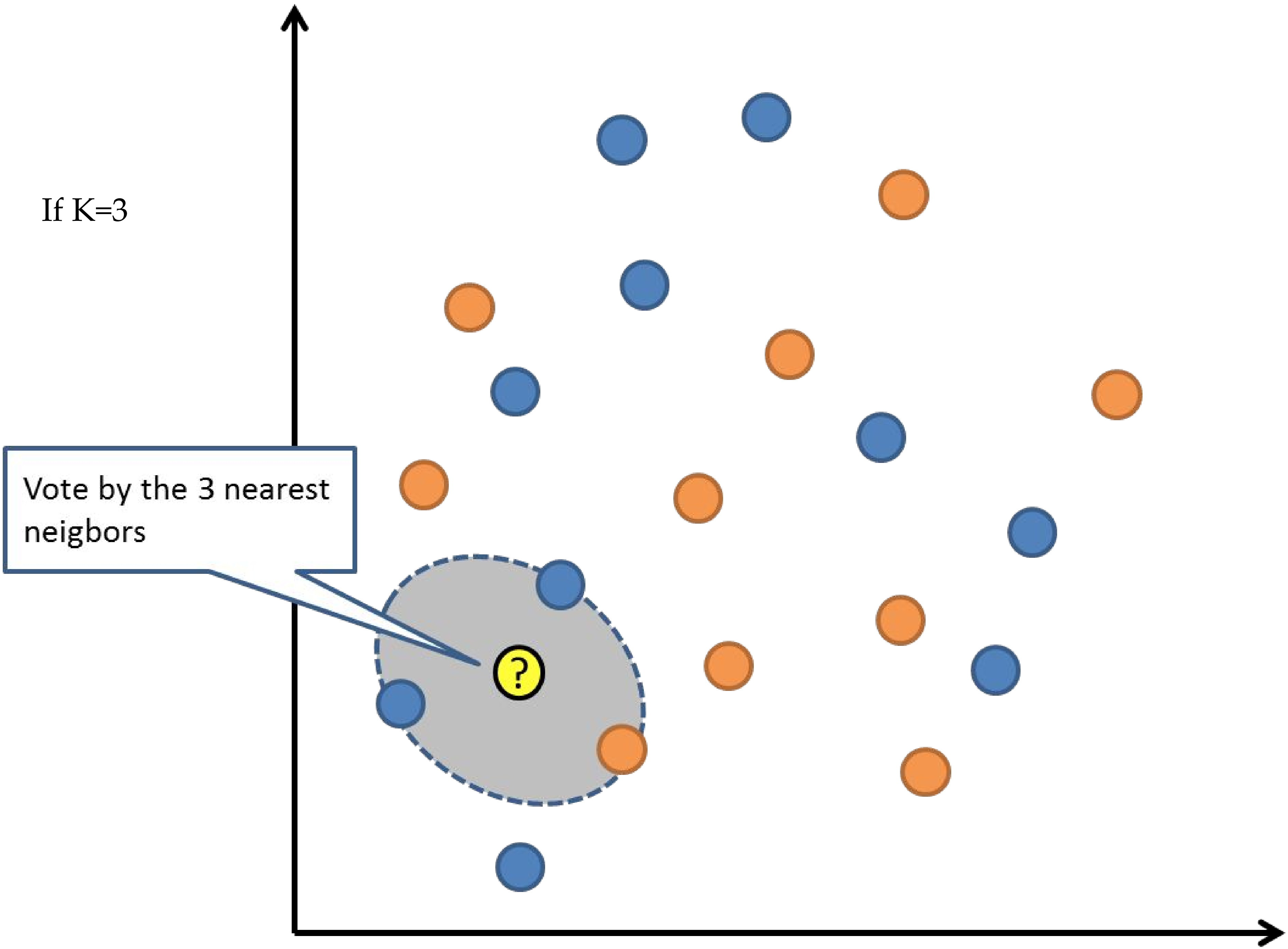


Supervised Learning Algorithm Example

K nearest Neighbors.

The input of the algorithm is a training corpus and a «K».

Given a test instance, the algorithm looks at the labels of the «K» nearest neighbors of the instance. The most frequent label of its neighborhood will be assigned to the test instance.


If K=3
Vote by the 3 nearest neigbors
?

To get the distance between two instances, we need to use a distance metric

Some Distance Metrics:

Euclidean Distance

$$d_E(P,Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}.$$

Cosine Similarity

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2} \sqrt{\sum\limits_{i=1}^{n} B_i^2}}$$

Clustering Algorithm:

K-Means

Input: Instance Set and a «K» that indicates the number of clusters that we want to obtain.

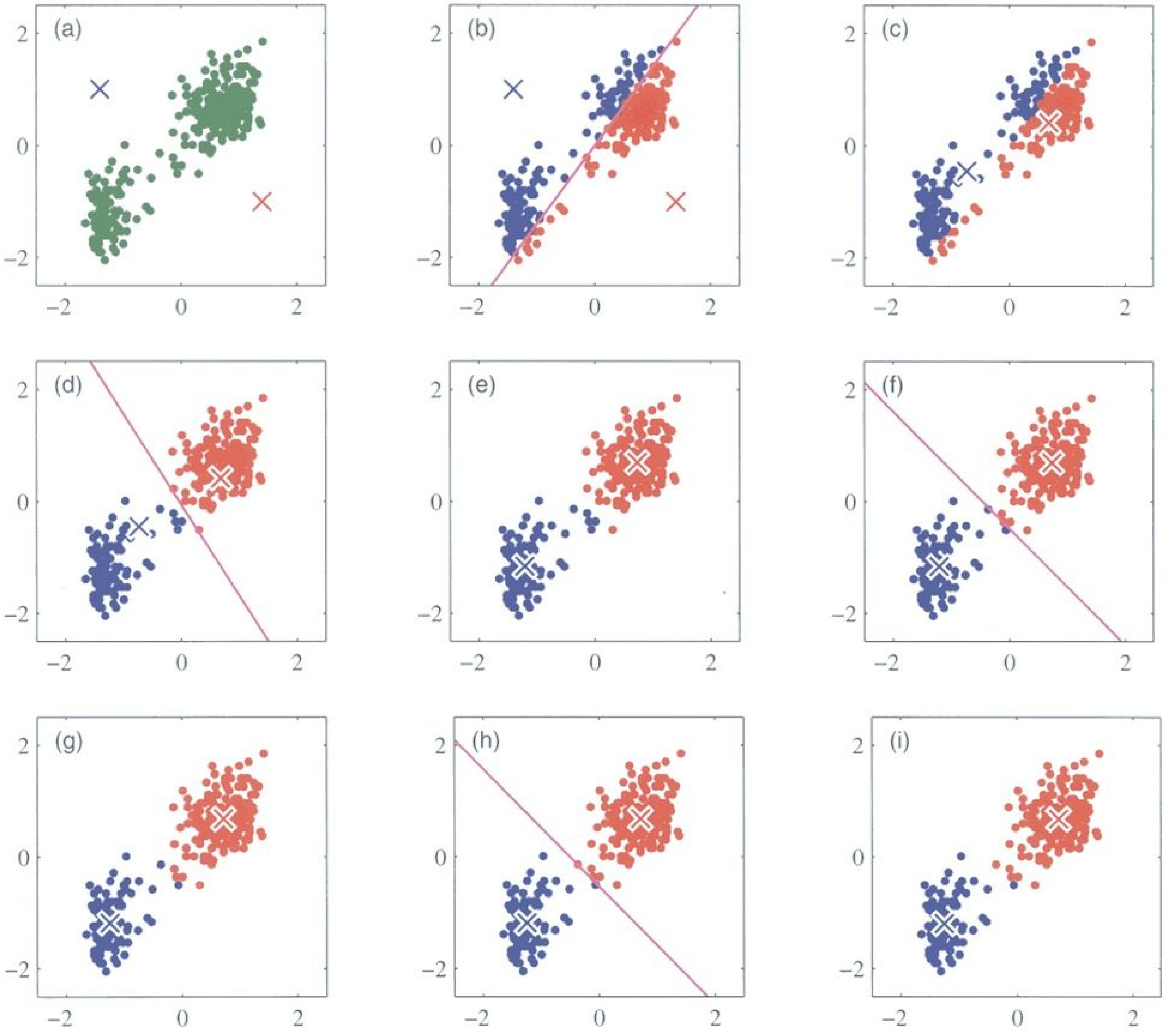
Initially, «K» elements are chosen at random (this is the simplest initialization strategy, much more complex ways improve drastically the precision of the algorithm). These elements will be called initial centroids.

Each instance will be assigned to its closest centroid.

The centroids are updated. This is done by calculating the mean by components of the assigned instances with that centroid.

If cluster 1 has vectors v1, v2 and v3 assigned, the updated centroid will be the mean by components of vectors v1,v2,v3.

The algorithm stops when there is an iteration in which the centroids dont change.


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
2
0
−2
−2
0
2

http://simplystatistics.org/2014/02/18/k-means-clustering-in-a-gif/kmeans/

What kind of features do I have to extract?

This will depend entirely on the problem at hand. The goal of these features is to characterize the instances in a way that these features can differentiate the instances with a label from the instances with a different label.

If every instance, regardless of the label, has the same value for one of the features, this feature will be completely useless.

To determine which features have to be extracted, the problem needs to be analyzed thoroughly.

Ex:

We want to classify texts by the age of the writer. (10year-olds vs 30year-olds).

Orthographic mistakes are more common in younger people.

Topics (10yo -> school, playing, their parents; 30yo -> mortgage, boyfriend/girlfriend, sadness).

Discourse complexity. Syntactic trees will be simpler in the texts of younger people. Shorter words. Shorter sentences. Tendency of not using subordination/coordination in their sentences. Less usage of the passive voice.

Vocabulary Richness.

Usage of exclamation marks.

What kind of curse words are used, if any?

…

Ex2:

We want to decide if a tweet is ironic or not.

Useful Features:

Unbalance between word frequency:

Very frequent words followed by a very uncommon one.

“Rajoy es un presidente del gobierno fantabuloso”.

“Rajoy es un presidente del gobierno correcto”.

First sentence, more likely to be ironic.

Number of senses that a word of the tweet has. If a word has a lot of senses, it is probable that it is being used in an ambiguous way (play on words, irony statement…).

Usage of curse words. Usage of certain characters such as exclamation/interrogation marks.

Tweet Sentiment.

More info and other features:

http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/download/5225/3029


HEY! YOU STILL ALIVE?
memegenerator.net

NLP Example Studies

- What do suspects’ words really mean?
- Author Profiling
- Alzheimer’s Detection
- Hate Speech

What do suspects’ words really mean?

- 1996 FBI paper.
  - https://www2.fbi.gov/publications/leb/1996/oct964.txt
- Analyzes real FBI cases. They explain how to analyze questioning transcriptions.
- Given texts that are questioning transcriptions, extract features to determine:
  - Is the suspect telling the truth?
  - Is he guilty?
  - Is he sure of what he is saying?

A case that shocked America.

The son and daughter of Susan Smith disappear. Susan tells the police that an african american kidnapped their children. She tells them he had a gun.

Both are devastated. They are interviewed on the news, crying.

They said to the press:

Susan: “My children wanted me. They needed me. And now I can’t help them”.

David: “They’re okay. They’re going to be back home soon”.

Really sad.

Poor Smiths.

UNION, S.C. (WCIV) -- It was 20 years ago on Saturday the small town of Union was thrust in the national spotlight for all the wrong reasons. A 23-year-old mother cried before a nation, claiming she had been carjacked and her two boys were gone.

Susan Smith's tears may have been real, but her story was not.

For nine long days she recounted the tale, describing a black man she said committed the crime.

"They made me feel like kind of bad coming to me up off my job early in the morning," an African American man told ABC New 4 at the time.

Fear fueled racial tensions in the tiny town. A kidnapper had yet to be caught. Three-year-old Michael and 14-month-old Alex were gone.

But what happened next shook the community to it's core.

Susan Smith confessed. She had strapped the boys in their seats and rolled the family car down the ramp of the local John D. Long Lake, leaving the boys to drown.

Susan: “My children wanted me. They needed me. And now I can’t help them”.

David: “They’re okay. They’re going to be back home soon”.

Statement Analisis

Given questioning transcriptions, extract the most information possible by analyzing the words that are used.

In this paper, this process is done manually.

What kind of features are extracted?

- Part of Speech:
    - Pronouns
    - Names
    - Verbs
- Length Metrics

Pronouns

They analyze the pronoun «I», «We» and possessive pronouns.

“I”

In real confessions, they detected that when a real story is told, the proun «I» is used frequently. A deviation from this can be considered suspicious.

“I got up at 7:00 when my alarm went off. I took a shower and got dressed. Met a man. Talked with him for a few minutes. I drove to work”.

“I” vs “We”

Using “I” indicates full responsibility. «We» can mean that the suspect is trying to avoid full responsibility.

“We”

Specially relevant in marriages. A lack of usage of this pronoun can indicate distance between the members of the couple.

In the cases in which there was an alleged rape or kidnapping:

“He forced me into the woods” standard statement.

“We went into the woods” kind of statement that is usually found in fake accusation cases.

In this case, «we» indicates a proximity between assaultant and victim that is not usual.

Possesive Pronouns (my, our, your, his, her …)

They emphasize the property of an object. Changing «My house to «the house» can be relevant.

In the article they have an example of the statements given by a guy that burnt his house down to get the insurance money in which he changes «my house» to «the house» when narrating the part when it burnt down.

Nouns

"… I lost control of the gun. I sensed that the barrel was pointing in Louise's direction …"

Statements in which the suspect changed «my wife» to Louise, without even using her name before. This distantiates himself from the fact that «Louise» was his wife.

Verbs

Verbal tense is very relevant. In the Smith case, the usage of past tense when talking about their dissapeared children is not standard. Generally the parents of kidnapped children maintain hope until the end and as a result, use the present tense to talk about their children.

Length Metrics

The extension of the statements of what happened before/after and during the event. (Before/after during the murder/kidnapping/assault...)

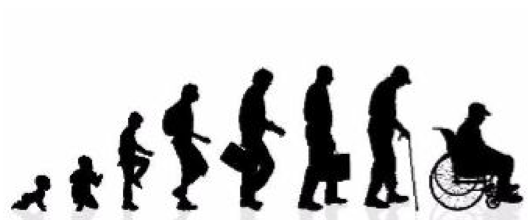
Non-convincing terms

The usage of doubt-filled expressions such as «kind of», «i mean» or «you know» make the statements be less credible.

Very interesting paper. If you have the time and motivation, read it.

Author Profiling

Author Profiling



Author Profiling

How can we do this?

1- Decide the labels, what are we going to classify. Ex: man vs woman, 10s vs 30s, engineer vs linguist…

2- Think about what features can be useful to differentiate between the labels.

3- Implement.

4- Machine Learning.

Gender Identification

We have texts in which we know if the author is a man or a woman.

What features can be extracted to differentiate between genders?

Gender Identification

2 types of features can be extracted

Content-Based

- Word frequency, N-gram frequency.

Effective approach, but domain-dependent and not very scalable.

Structure-Based

- Analyze the structure of the text (syntactic/discursive…)

Domain Independent. Non-trivial features.

Gender Identification

Character-based

Word-based

Sentence-based

Syntactic

Sentiment-based

Gender Identification

Character-based

Ex:

% of the characters of a text that are commas, dots, colons, semi-colons, hyphens, quotations, parenthesis, interrogation/exclamation marks, emotes…

Word-based

Ex:

AVG characters per word, vocabulary richness, usage of acronyms, % of the words that are pronouns, adjectives, verbs, stopwords…

Gender Identification

Sentence-based

Ex:

Avg number of words per sentence, difference between the longest and shortest sentence…

Sentiment-based

Ex:

% of the words that are positive/negative, usage of abbreviations, curse words…

Gender Identification

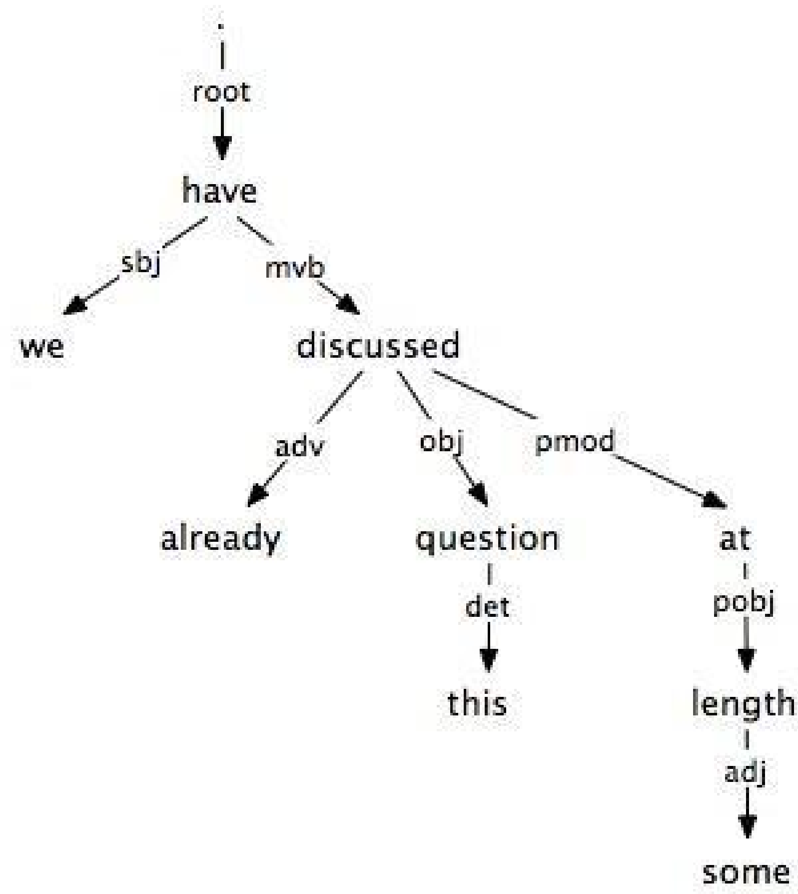
Syntactic

Ex:

Width and Depth of the syntactic trees. It is an indication of the complexity of the discourse.

Number of different syntactic dependencies used per sentence.

Percentage of dependencies that are «obj», «pmod»…



Feature	Male	Female
Mean # words	1221.54	1242.91
Mean chars per word	5.071	5.074
Mean num of chars	7465.15	7576.29
% uppers	3.088%	2.846%
mean different words	459.67	601.24
mean #commas	50.808	74.567
mean #dots	42.95	67.25
mean #exclamation	0.413	0.586
mean #questionmarks	2.743	3.034
mean #sentences	31.696	44.756
mean wordspersentence	26.0362	28.1192
mean #stopwords	337.2598	460.431
mean acronyms	1.4234	1.4629
mean positive words	21.033	26.011
mean negative words	20.601	28.923
mean political words	15.7758	11.8876

We have all these features, now what?

For each text, we will have a feature vector like this:

[feat1, feat2, feat3 …. featN]

N-dimensional vectors labeled with the correct label. This will be a Supervised Learning Problem.

A % of the instances will be the training set, the instances that the classifier will use to «learn».

The rest will be used as test set, the classifier will predict their labels and these predictions will be compared with the real values to compute the accuracy of the system.

Features Used	Accuracy AuthorshipDat	Accuracy LiteraryAmerican
Full Set	**89,97%**	**90,71%**
Character-based	87.91%	81.02%
Word-based	81.18%	78.79%
Sentence-based	65.01%	73.88%
Dictionary-based	71.45%	84.39%
Syntactic	85.17%	90.76%
Discourse	75.34%	75.22%
Function Words (FW)	81.72%	52.73%
Stopwords (SW)	81.46%	52.73%
Parts of Speech (PoS)	81.53%	74.84%
FW + PoS	82.67%	76.81%
SW + PoS	82.88%	76.87%

Table 4.8: Results of the gender identification experiments on both datasets.

	English	Spanish	German	French	Catalan	Italian
Accuracy	80.24%	88.02%	77.87%	83.98%	88.11%	86.54%
MajClassBaseline	50%	50%	50%	50%	50%	50%
Number of Features	96	83	73	52	79	52

Which features work better?

Women tend to have:

- Sentences that are more complex syntactically than men’s
- More usage of negative/positive words.
- More vocabulary richness.
- More tendency of using words describing «non-tangible» concepts.

Men tend to use:

- Shorter sentences.
- «Tangible» words.
- More «action» description instead of «thought» description.
- More word repetition.

1- Know what kind of texts you have as input.

- Journalistic? Tweets? Stephen King novels?

Each kind of text has its own structure an inner characteristics.

2- Think about what kind of features can be used to describe each instance and differentiate between the instances of other kind. This is the most important part.

Ex: If we are analyzing tweets: number of hashtags, links, if the link is from instagram, if we have exclamations, if there is a picture on the tweet, mentions, usage of common or uncommon words, orthography mistakes…

3- Decide if it is a supervised, semisupervised or unsupervised learning problem.

4- Choose a ML algorithm and apply. Analyze the results.

Alzheimer's Detection

upf. taln

ALZHEIMER'S DISEASE IS THE 6TH LEADING CAUSE OF DEATH IN THE UNITED STATES

In 2017, Alzheimer's and other dementias will cost the nation $259 billion
By 2050, these costs could rise as high as
$1.1 TRILLION

MORE THAN 5 MILLION AMERICANS ARE LIVING WITH ALZHEIMER'S
BY 2050, THIS NUMBER COULD RISE AS HIGH AS 16 MILLION

EVERY 66 SECONDS
someone in the United States develops the disease

35% of caregivers for people with Alzheimer's or another dementia report that their health has gotten worse due to care responsibilities, compared to **19%** of caregivers for older people without dementia

1 IN 3
seniors dies with Alzheimer's or another dementia

Since 2000, deaths from heart disease have decreased by 14%
while deaths from Alzheimer's disease have increased by 89%

MORE THAN **15 MILLION AMERICANS** provide unpaid care for people with Alzheimer's or other dementias

IN 2016 these caregivers provided an estimated **18.2 BILLION HOURS** of care valued at over **$230 BILLION**

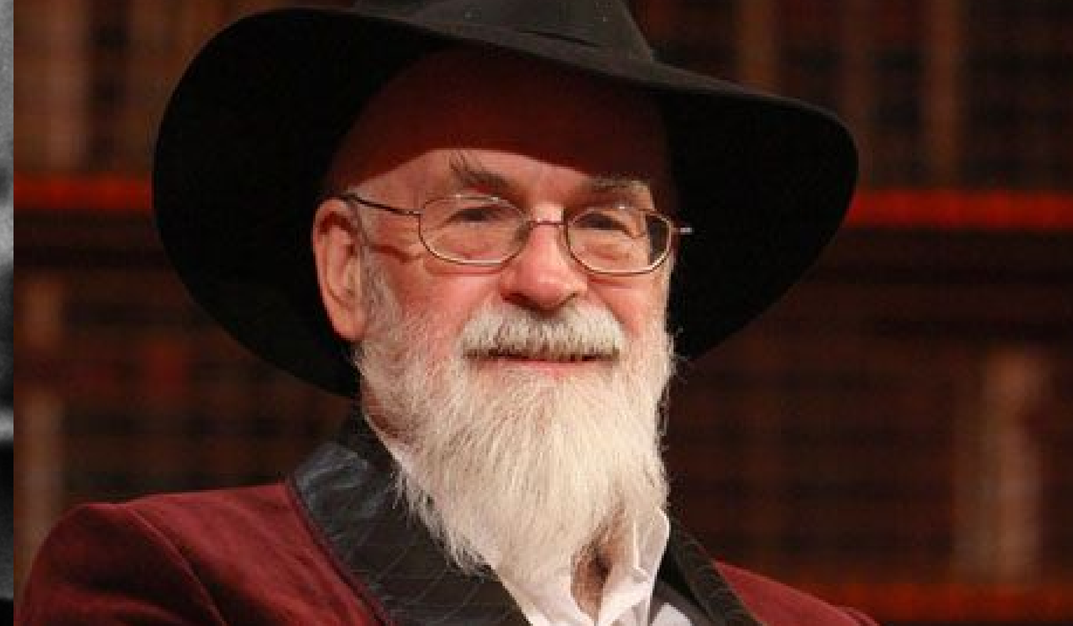
IT KILLS MORE THAN
breast cancer and prostate cancer
COMBINED

What can we do?

The characteristic symptoms of Alzheimer's are difficulties with memory, language, problem solving and other cognitive skills that affect a person's ability to perform, for instance, everyday activities.

People with Alzheimer's have trouble in following conversations, struggle with vocabulary and have problems to express themselves with precision.

From the natural language processing point of view, the effects of Alzheimer's can be assessed through the analysis of the writing style of an author before and after the break-out of the disease.

3 Authors that wrote some of their novels under the influence of Alzheimer's disease.

Agatha Christie
Terry Pratchett
Iris Murdoch

We select books pre and post disease. 1 IM pre 1 IM post, 3 AC pre 3 AC post, 4 TP pre 4 TP post.

From these texts we extract:

- Character-based Features
- Word-based Features
- Sentence-based Features
- Dictionary-based Features
- Syntactic
  - PoS, Dependencies, Shape
- Lexical

Experiments

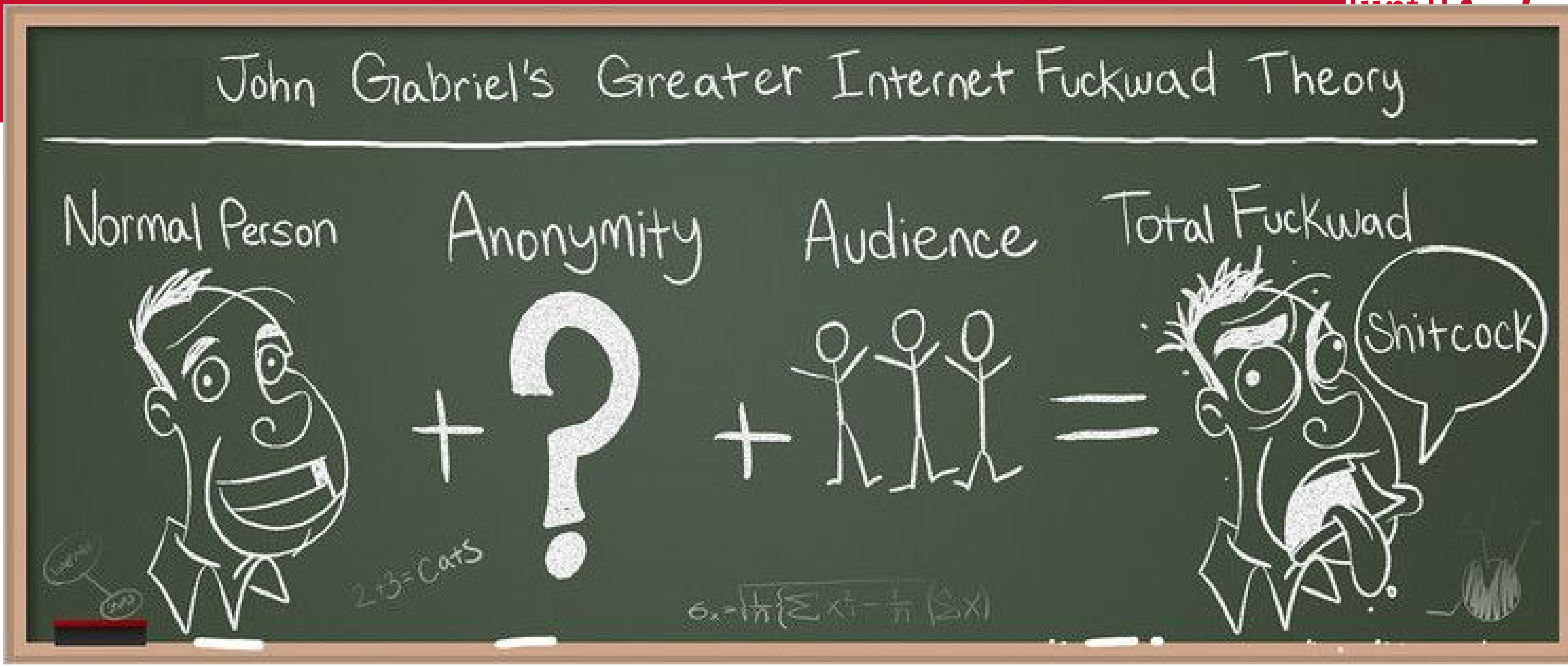
Features Used	Author Id	Alzheimer's Id	Book Id
Full Set	**96,39%**	**82,21%**	**73,02%**
Char	69,75%	64,91%	35,04%
Word	83,44%	70,60%	42,65%
Sent	61,65%	60,85%	19,25%
Dict	71,58%	65,56%	33,33%
Syntactic	94,71%	73,83%	55,15%
Lexical	57,45%	54,47%	19,98%
Majority Class	50%	50%	6%
Token 2-gram 100	80,23%	68,44%	33,27%
Token 2-gram 300	84,15%	72,52%	40,39%
Token 2-gram 500	85,87%	74,60%	48,06%
Token 2-gram 700	89,47%	76,66%	52,94%
Token 2-gram 900	90,87%	78,01%	57,35%

Experiments

Features Used	Iris Murdoch	Agatha Christie	Terry Pratchett
Full Set	98,50%	86,00%	94,50%
Char	82,33%	72,67%	76,63%
Word	87,17%	68,51%	85,00%
Sent	68,67%	59,33%	69,29%
Dict	74,50%	66,50%	76,50%
Syntactic	89,51%	76,89%	86,21%
Lexical	75,83%	71,61%	67,13%
Majority Class	50%	50%	50%
Token 2-gram 100	85,67%	64,50%	83,71%
Token 2-gram 300	87,01%	65,17%	85,63%
Token 2-gram 500	88,55%	65,94%	85,33%
Token 2-gram 700	90,19%	65,39%	87,95%
Token 2-gram 900	90,66%	69,17%	88,78%

Hate Speech Detection

upf. taln


John Gabriel's Greater Internet Fuckwad Theory
Normal Person
Anonymity
Audience
Total Fuckwad
Shitcock
2+3=Cats


ON THE INTERNET
nobody knows you're a dog

Hate Speech

Hate Speech

“Hate speech can be defined as speech designed to promote hatred on the basis of race, religion, ethnicity or national origin.

Hate speech is directed to a specific target group, stigmatizes the target group by implicitly or explicitly assigning it qualities widely regarded as highly undesirable and influences people to view the target group as undesirable and a legitimate object of hostility.”

EXAMPLE:

1994’s Rwandan Genocide

- Hate speech disseminated by radio.
- 75% of the tutsi population was killed.

MINORITIES

56%

of Europeans think that discrimination on grounds of ethnic origin is widespread

European Commission (2012) Discrimination in the EU in 2012

21%

of survey respondents had personally experienced at least one incident of anti-semitic verbal insult or harassment, and/or a physical attack in the past 12 months

GENDER

Up to 21% of young women have received unwanted sexually explicit emails or text messages

Up to 28% have been the target of offensive propositioning on social networking sites or internet chat rooms

All from EU Agency for Fundamental Rights (2013) Report of violence against women

SEXUAL ORIENTATION

1/4

In the last five years, a quarter of all respondents to a survey of LGBT people said they had been attacked or threatened with violence because of their sexuality, with almost half reporting discrimination*

1/5

Up to 1 in 5 of the 93,000 LGBT people surveyed across the EU said that their last harassment was online*

15%

As many as 15% said their most serious incidence of harassment was on the internet*

50%

Close to half of Europeans believe that discrimination on grounds of sexual orientation is widespread in their country**

** EU Agency for Fundamental Rights (2013) EU LGBT survey*
***European Commission (2012) Discrimination in the EU in 2012*

HATE INCIDENT MOTIVATION

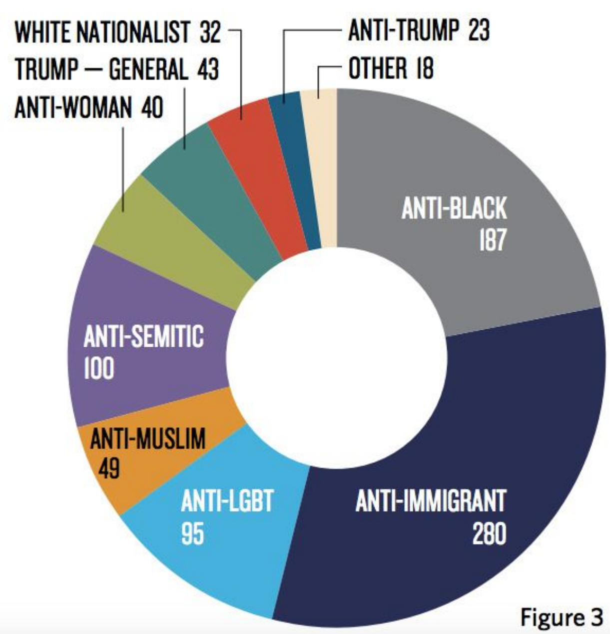
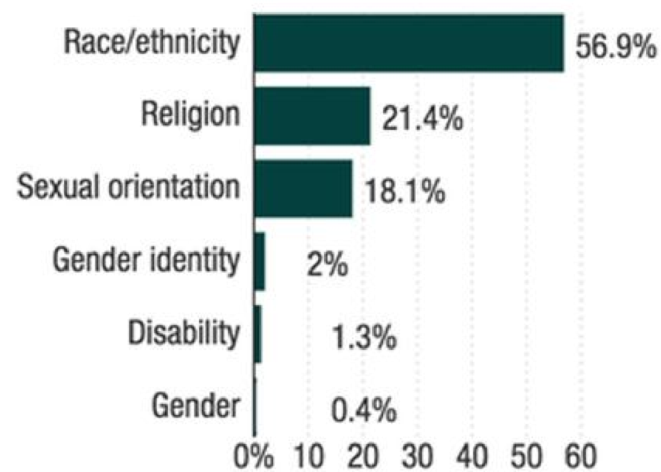


Figure 3

Motivations for hate crime incidents, 2015



The corpus

Some corpora is available, but mostly Twitter-based.

Forums and blog posts, not considered in the related work.

OffSet -> New corpus of **offensive/abusive speech.**

The corpus

Category	Sources	#instances
Racism	v/coontown	46
	v/coonshire	81
	v/niggers	11
	shitskin	41
	niggermania	49
	chimpmania	107
	debate	2
	Total	**337**
Misogyny	returnofkings	100
	rooshv	115
	r/redpill	2
	debate	1
	r/womenhate	4
	misogynyforum	8
	Total	**230**
Antisemitism	stormfront	289
	debate	1
	Total	**290**
fatshaming	r/fatpeoplestories	62
	v/fatpeoplehate	225
	Total	**287**
antilgbt	godhatesfags	44
	returnofkings	7
	breitbart	141
	debate	3
	Total	**195**
notoffensive	stormfront	3
	rooshv	13
	returnofkings	16
	serenesforest	200
	debate	173
	Total	**405**
TOTAL		**1744**

	All classes	Racism	Misogyny	Antisemitism	Fatshaming	Antilgbt	notoffensive
Mean number of tokens	339.54	144.95	682.13	376.66	321.71	341.93	291.80
Min number of tokens	7	8	36	18	20	7	70
Max number of tokens	3654	1533	3654	2915	2999	2871	1700
Mean token length	4.62	4.58	4.57	4.71	4.40	4.82	4.70
Mean number of chars	1915	812.56	3851.51	2150.96	1715.97	1977.92	1677.70
Min number of chars	51	52	189	76	106	51	412
Max number of chars	20007	8554	20007	16482	15725	16770	10918
Mean number of sentences	18.16	8.66	33.85	20.43	19.82	17.26	14.79

The sources were crawled and only offensive texts were selected, off-topic and non-relevant texts were discarded.

Whenever there was any doubt that a text was offensive or not, it was discarded.

Racism:
“Some niggers are such unruly creatures that in order to make them anything other than criminals you must first strip away their agency and refer to them as if they were objects being acted upon by nebulous external forces.”

Misogyny:
“No women do not scare me. They are pathetic manipulators, nothing more. Men can be better manipulators than them because they are capable of entering the female frame of mind, yet most do not do so due to social stigma.”

Antisemitism:
“Jews are obvious very intelligent, but present intelligence in everything what is bad, evil, twisted and pervert!!! So they are smart evil race on the earth! There will be one moment or we (Aryans) or Jews, because obvious both can’t live on one planet”

Fatshaming:
“Super obese ham planet gets so big that only super obese ham planets will spend time with it. Blabs on and on about how they’re beautiful on the outside and inside, and how society is prejudiced and immoral for judging them.”

Antilgbt:
“Homosexuality and legalized same sex marriage has spread across the U.S. infecting the masses, and you love to have it so. But take heart, the real apocalypse is coming soon to a graveyard near you, where the bodies will be stacked liked cords of wood.

Relevance of the sources

- More than 100 murders are linked with users of **Stormfront**. Often called “The murder capital of the internet”. Don Black, the guy behind Stormfront was a former Ku Klux Klan leader. Anders Breivik, responsible for killing 77 people in Norway in 2011.

- Return of kings and Rooshv promote sexism and contributes to the discrimination and objectification of women.

- GodHatesFags is a site from the “Westboro Baptist Church” which promotes hatred towards the LGBT collective.

...

Experiments

Benchmark experiments to test different offensive speech classification techniques.

Feature sets:

- Simple bag-of-words approach
- Author Profiling features
  - Character-, word-, sentence-, dictionary-based, syntactic and discourse features.
- Google news embeddings
  - Each document is represented by a vector with the mean of the embeddings of each word.
- Offensive embeddings
  - Every thread of Stormfront and every blog post of Return of Kings and Rooshv were used to train word embeddings (every text apart from the ones used in the OffSet corpus). Used the same way as the Google embeddings.

Experiments

Supervised Machine learning.

10-fold cross-validation with LibSVM with a linear kernel (Weka’s implementation)

3 kinds of experiments:

- Binary (offensive vs not offensive)
- Multiclass (racism, misogyny, antilgbt, antisemitism, fat shaming, non-offensive)
- One-vs-rest (racism vs non-racism, misogyny vs non-misogyny…)

Results

Approach	F-measure binary	F-measure multiclass
SVM with Stylistic features	82.45%	62.39%
SVM with Google embeddings	**87.56%**	**80.33%**
SVM with Offensive embeddings	86,56%	76,49%
SVM with BoW100	78.44%	55.21%
SVM with BoW300	82.11%	67.26%
SVM with BoW500	84.46%	69.72%
SVM with BoW700	83.42%	70.41%
SVM with BoW900	83.14%	71.61%
Majority Class	76.78%	23.22%

Approach	Antisemitism vs. rest	Misogyny vs. rest	Racism vs. rest	Antilgbt vs. rest	Fatshaming vs. rest
Stylistic features	85.32%	93.06%	85.21%	89.91%	88.81%
Google embeddings	92.14%	94.38%	92.02%	**94.32%**	**96.27%**
Offensive embeddings	**92.43%**	**95.02%**	86.64%	94.15%	93.11%
BoW100	86.46%	91.45%	80.73%	88.81%	86.12%
BoW300	90.48%	93.29%	**92.37%**	91.68%	90.76%
BoW500	90.77%	93.29%	90.94%	92.49%	91.06%
BoW700	91.28%	93.01%	90.65%	92.49%	91.04%
BoW900	91.39%	92.61%	89.85%	93.06%	91.80%
Majority Class	83.37%	86.81%	80.68%	88.81%	83.54%

Resources

Everything is publicly available either at:

https://github.com/joanSolCom/Datasets

or at a drive folder that I have public:

https://drive.google.com/drive/folders/1_Zz_E3G140RqMYkF0A29E41Q3NP_dALH?usp=sharing


That's all Folks!