ntic
master
revolucionamos la comunicación

UNIVERSIDAD
COMPLUTENSE
DE MADRID

Deep Learning

Alberto Ezpondaburu

Analytical Index

1. Introduction to Deep Learning
2. Convolutional Neural Networks (CNN)
3. Recurrent Neural Networks (RNN)
4. Natural Language Processing with Deep Learning
   a. Introduction to NLP
   b. Word Embeddings
   c. Attention and Contextual Word Representation
5. Deep Generative Modelling

4

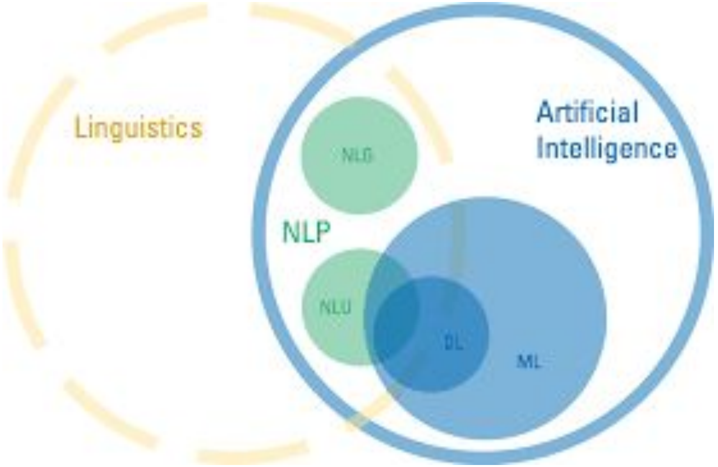
Natural Language Processing With Deep Learning

4.1

Introduction to NLP

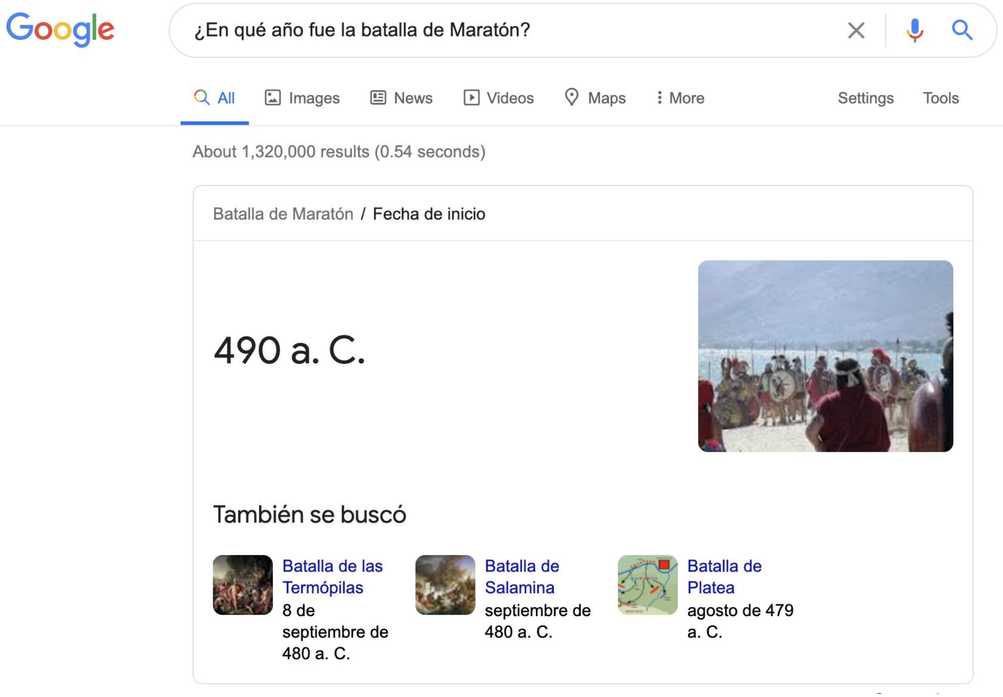
What is NLP?

- Natural Language Processing (NLP)
- Intersection between AI, linguistics and computer science
- You almost always work with text, but also voice
- It's keyword matching, it's text normalization, it's ML, it's clustering ...
- Human-machine understanding



~~*"Every time I fire a linguist, the performance of the speech recognizer goes up"*~~
Frederick Jelinek (1932 – 2010)

What is NLP?



¿En qué año fue la batalla de Maratón?

About 1,320,000 results (0.54 seconds)

Batalla de Maratón / Fecha de inicio

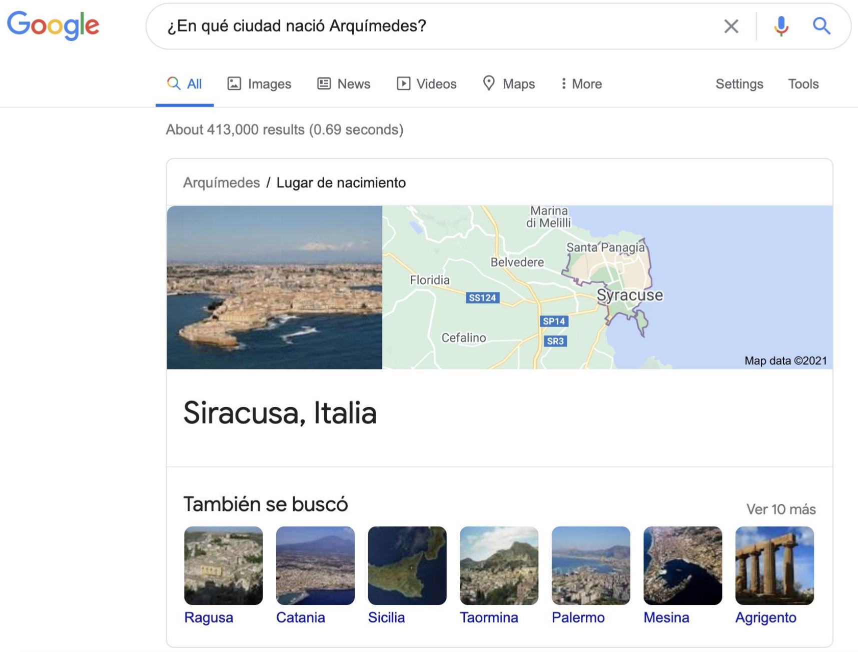
490 a. C.

También se buscó

Batalla de las Termópilas
8 de septiembre de 480 a. C.

Batalla de Salamina
septiembre de 480 a. C.

Batalla de Platea
agosto de 479 a. C.



¿En qué ciudad nació Arquímedes?

About 413,000 results (0.69 seconds)

Arquímedes / Lugar de nacimiento

Siracusa, Italia

También se buscó

Ver 10 más

Ragusa
Catania
Sicilia
Taormina
Palermo
Mesina
Agrigento

Natural Language Processing: Traditional NLP

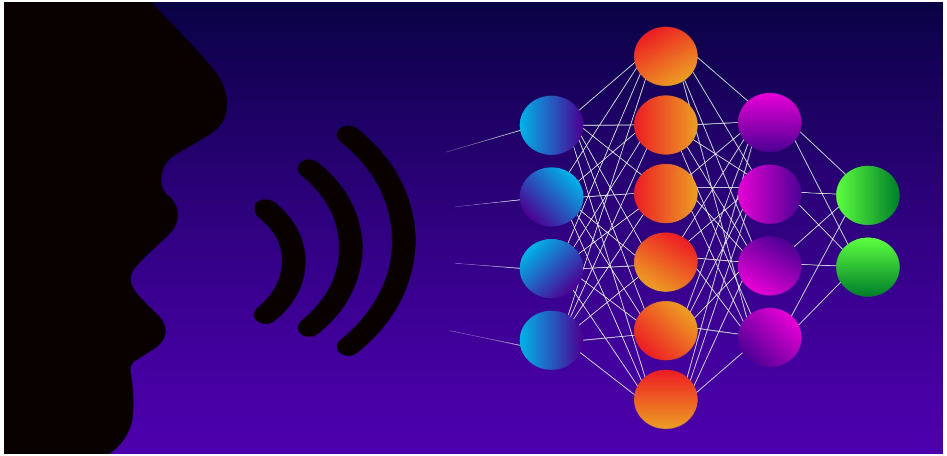
- Document classification/clustering
- Sentiment analysis in comments about brands, products, markets, politicians, ...
- Search for information in a search engine
- Detection of relevant topics in social networks

Natural Language Processing: Modern NLP

- **Conversational Agent, Dialogue System**:
  - Language input: **automatic speech recognition** and **natural language understanding**.
  - Language output (**natural language generation** and **speech synthesis**)

- **Machine Translation:** Automatically translate a document from one language to another

- **Web-based question answering:** Generalization of simple web search
  - What time does the store close?



Knowledge In NLP

In any ordinary data processing application, like bytes count, we don't need knowledge about what it means to be a word, in NLP...

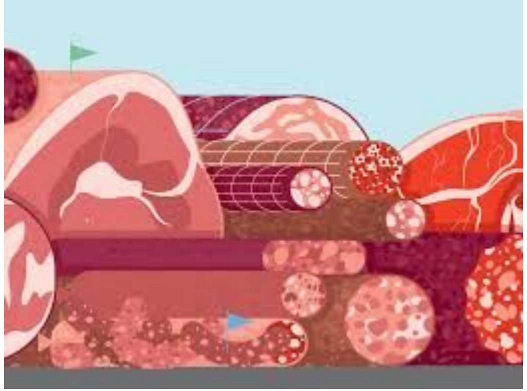
- **Phonetics and Phonology**: knowledge about linguistic sounds
  - Speech recognition and generation.
- **Morphology:** knowledge of the meaningful components of words
  - Casa, casas, casitas, casoplón...
- **Syntax**: knowledge of the structural relationships between words
  - Los alumnos harán los deberes. - Los deberes serán hechos por los alumnos.
- **Semantics:** knowledge of meaning
  - *"Zapatos de piel de señora"*
- **Pragmatics:** knowledge of the relationship of meaning to the goals and intentions of the speaker (Context)
  - *¿Tienes hora?* ... en la calle / en el médico
- **Discourse**: knowledge about linguistic units larger than a single utterance.
  - *¿Quién era el presidente* ***ese año****?*
- **World:**
  - ¿Cuántos países hay en Europa?

NLP: Ambiguity

El cura recibió una cura.

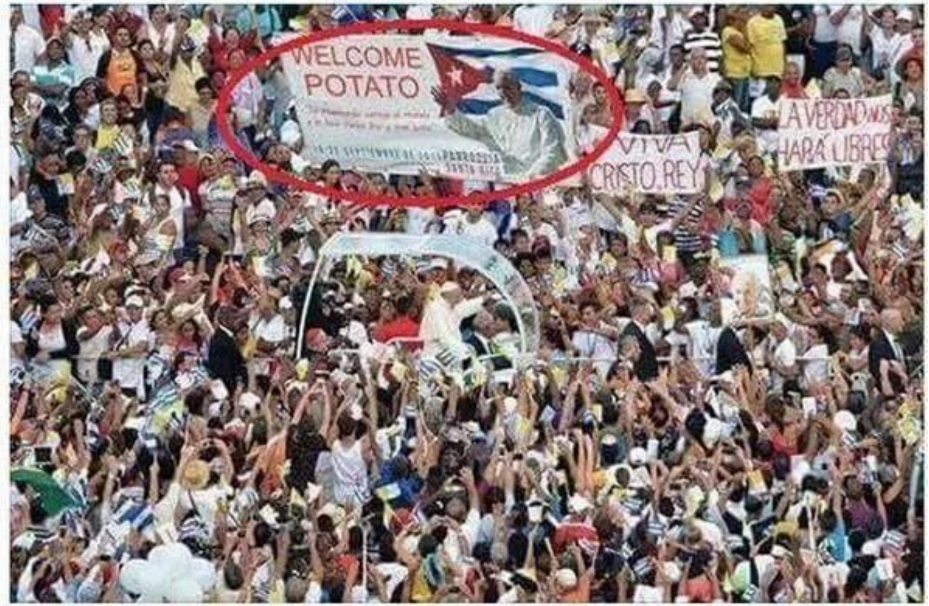
El policía encontró al sospechoso con unos prismáticos.

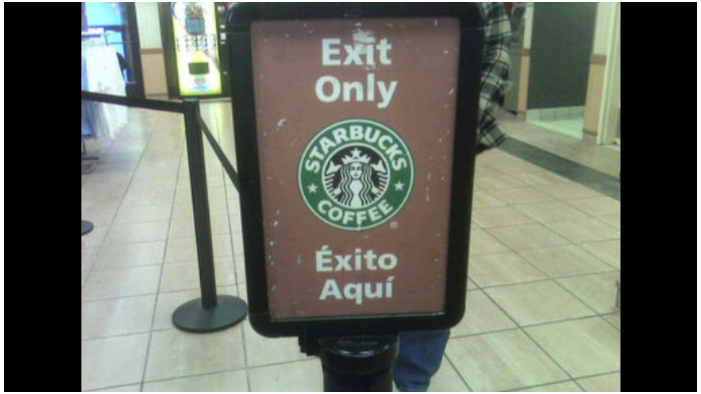
… por si las moscas



¡Vaya Chorizos!

NLP: Ambiguity





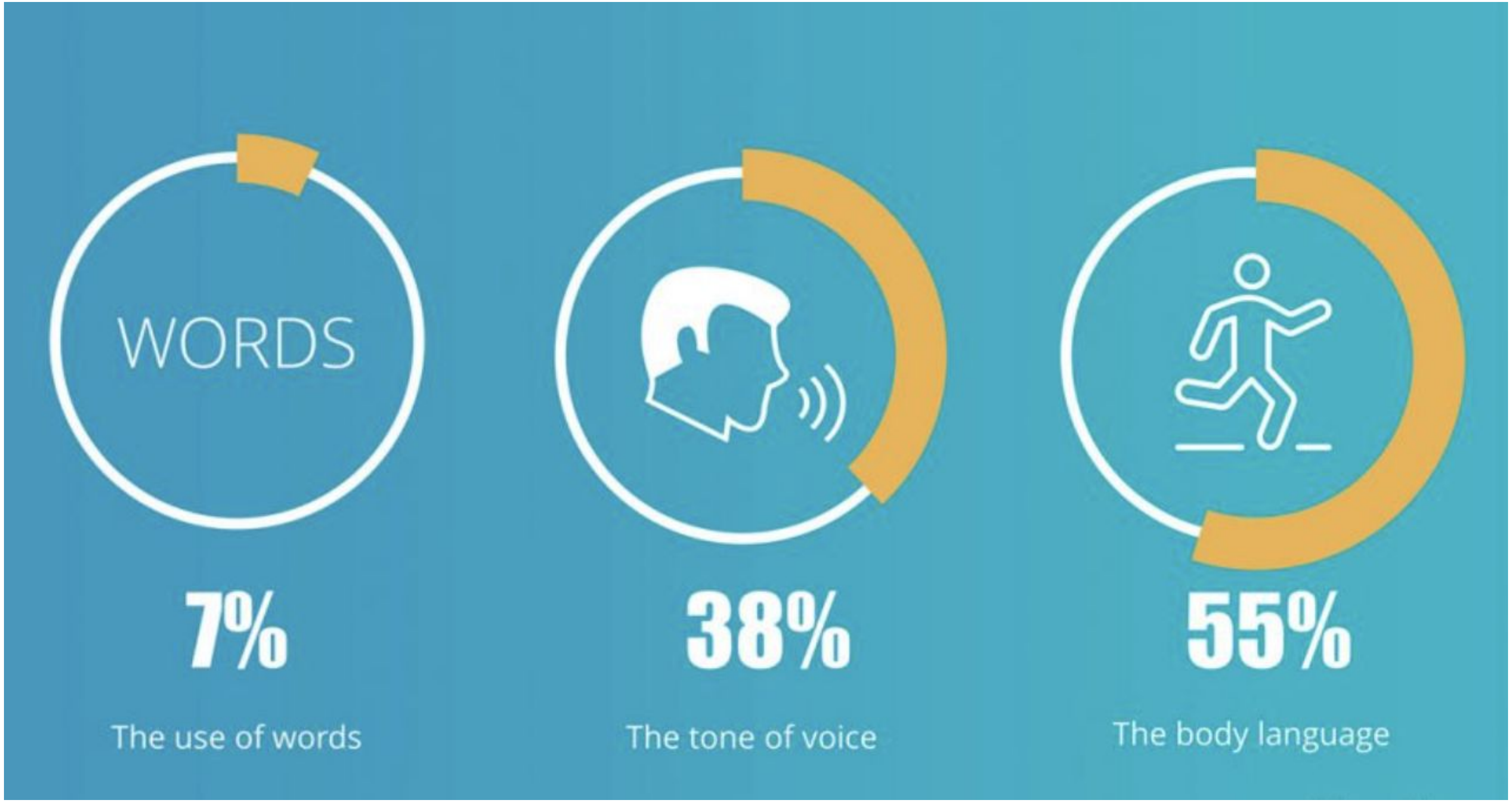
NLP: Problems

Entre lo que pienso,
lo que quiero decir,
lo que creo decir,
lo que digo,
lo que quieres oír,
lo que oyes,
lo que crees entender,
lo que quieres entender,
lo que entiendes

Existen 9 posibilidades de no entenderse

NLP: Problems

Mehrabian's Rule on non-verbal communication



4.2

Word Embeddings

Words as discrete symbols

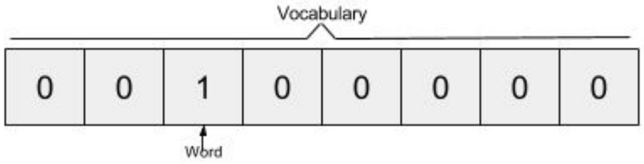
Words are processed as discrete symbols, |V| (length of vocabulary).

For ML models they are categorical variables.

One-hot encoding: Vectors with dimension |V| , all 0’s and one 1

house = [0 0 1 0 … 0]

home = [0 0 … 0 1 0]



Really **big** and **sparse** vectors (500 000 words).

All vector are orthogonal, How can we codify the **similarity** ?

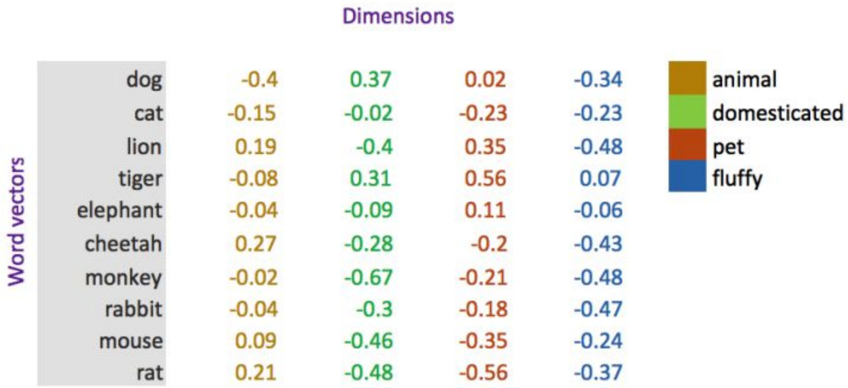
Distributional Semantics: words represented by their context

- **Distributional Hypothesis**: Similarity in meaning results in similarity of linguistic distribution. Words that are used and occur in the same contexts tend to purport similar meanings.
- "*A word is characterized by the company it keeps*" J. Firth, 1950s
- One of the most successful ideas in statistical NLP

Estoy **seguro** de que llegaré a tiempo.
¿Cuánto costará el **seguro** del coche?
Ese distrito no es **seguro**, es peligroso.

Word Vectors

- Words from the vocabulary are mapped to vectors of real numbers with lower dimension (50-300).
- Capturing distributional characteristics (relative meaning)=> measure similarity.



Word vectors	Dimensions			
dog	-0.4	0.37	0.02	-0.34
cat	-0.15	-0.02	-0.23	-0.23
lion	0.19	-0.4	0.35	-0.48
tiger	-0.08	0.31	0.56	0.07
elephant	-0.04	-0.09	0.11	-0.06
cheetah	0.27	-0.28	-0.2	-0.43
monkey	-0.02	-0.67	-0.21	-0.48
rabbit	-0.04	-0.3	-0.18	-0.47
mouse	0.09	-0.46	-0.35	-0.24
rat	0.21	-0.48	-0.56	-0.37

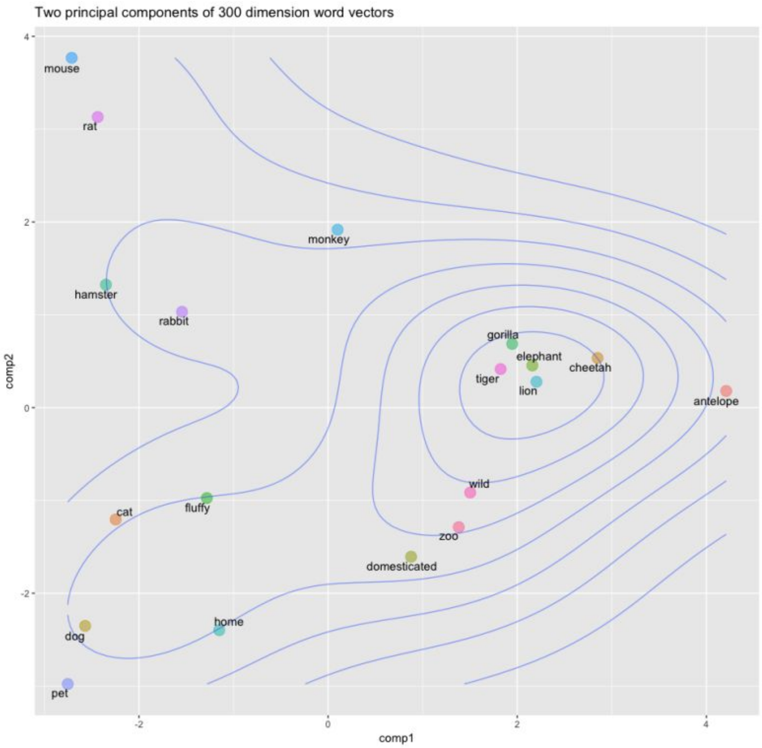
🟫 animal
🟩 domesticated
🟥 pet
🟦 fluffy

dog = [-0.4 0.37 0.02 -0.34]

cat = [-0.15 0.02 -0.23 -0.23]

https://medium.com/@jayeshbahire/introduction-to-word-vectors-ea1d4e4b84bf

Word Vector Space



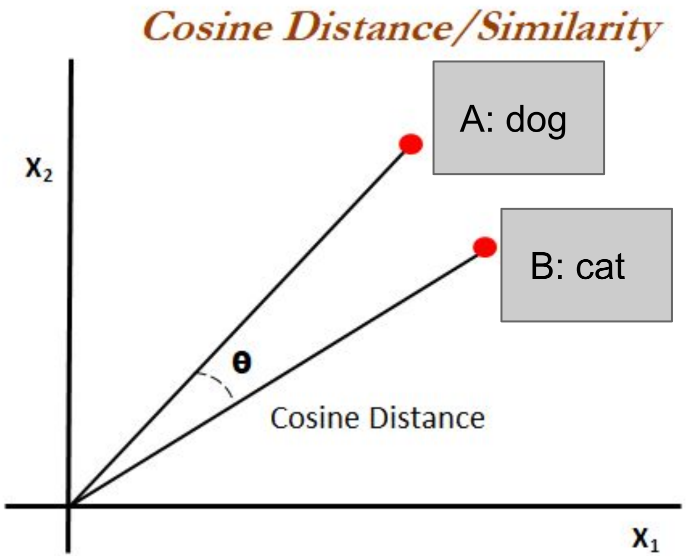
https://medium.com/@jayeshbahire/introduction-to-word-vectors-ea1d4e4b84bf

Document Based Vector

Count the **number of occurrences** of a word within a certain category

	ECONOMY	SPORTS	SCIENCE
Transfer	400	350	7
Speed	25	200	300

Similarity



- Euclidean distance:

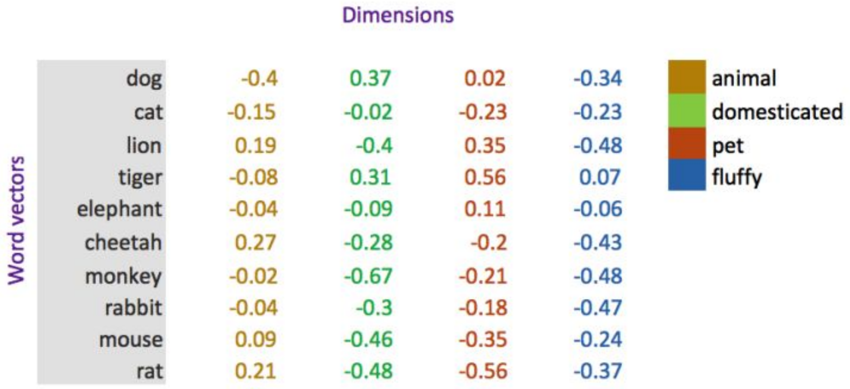
$$d(\overrightarrow{A}, \overrightarrow{B}) = \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2} = ||\overrightarrow{A} - \overrightarrow{B}||_2$$

- Cosine Similarity:
  - Cosine of the angle between the two vectors.

$$\overrightarrow{A} \cdot \overrightarrow{B} = ||\overrightarrow{A}|| ||\overrightarrow{B}|| \cos(\theta)$$

$$sim(A, B) = \cos(\theta) = \frac{\overrightarrow{A} \cdot \overrightarrow{B}}{||\overrightarrow{A}|| ||\overrightarrow{B}||}$$

Similarity



Dimensions

Word vectors				
dog	-0.4	0.37	0.02	-0.34
cat	-0.15	-0.02	-0.23	-0.23
lion	0.19	-0.4	0.35	-0.48
tiger	-0.08	0.31	0.56	0.07
elephant	-0.04	-0.09	0.11	-0.06
cheetah	0.27	-0.28	-0.2	-0.43
monkey	-0.02	-0.67	-0.21	-0.48
rabbit	-0.04	-0.3	-0.18	-0.47
mouse	0.09	-0.46	-0.35	-0.24
rat	0.21	-0.48	-0.56	-0.37

🟫 animal

🟩 domesticated

🟥 pet

🟦 fluffy

dog = [-0.4 0.37 0.02 -0.34]

cat = [-0.15 0.02 -0.23 -0.23]

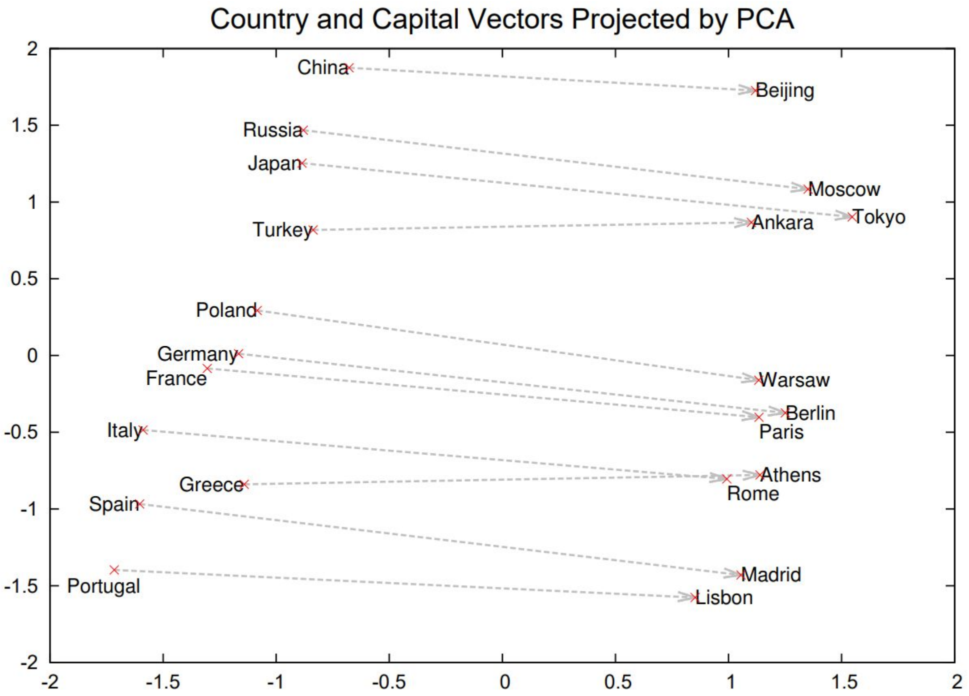
cheetah = [0.27 -0.28 -0.2 -0.43]

$$sim(A,B) = \cos(\theta) = \frac{\vec{A} \cdot \vec{B}}{||\vec{A}||||\vec{B}||}$$

sim(dog, cat) => -0.4 * (-0.15) + 0.37*0.02 + 0.02 * (-0.23) -0.34*(-0.23) = +0.141

sim(dog, cheetah) => -0.4 * 0.27+ 0.37*(-0.28) + 0.02 * (-0.2) -0.34*(-0.43) = -0.26

Word vectors operations



Find most similar words:

school: teacher, student, classroom

Find patterns:

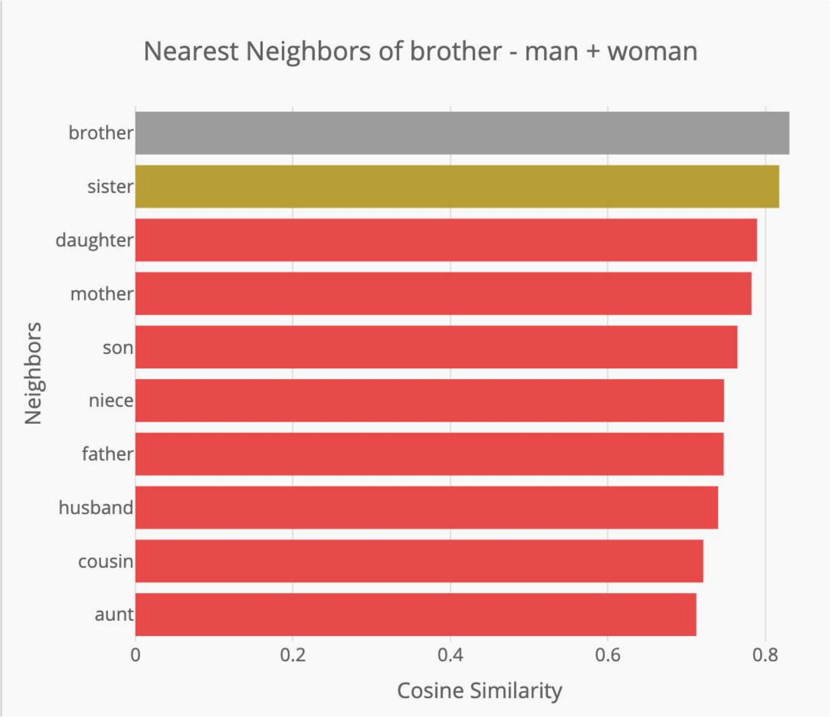
spain - madrid ≅ france - paris

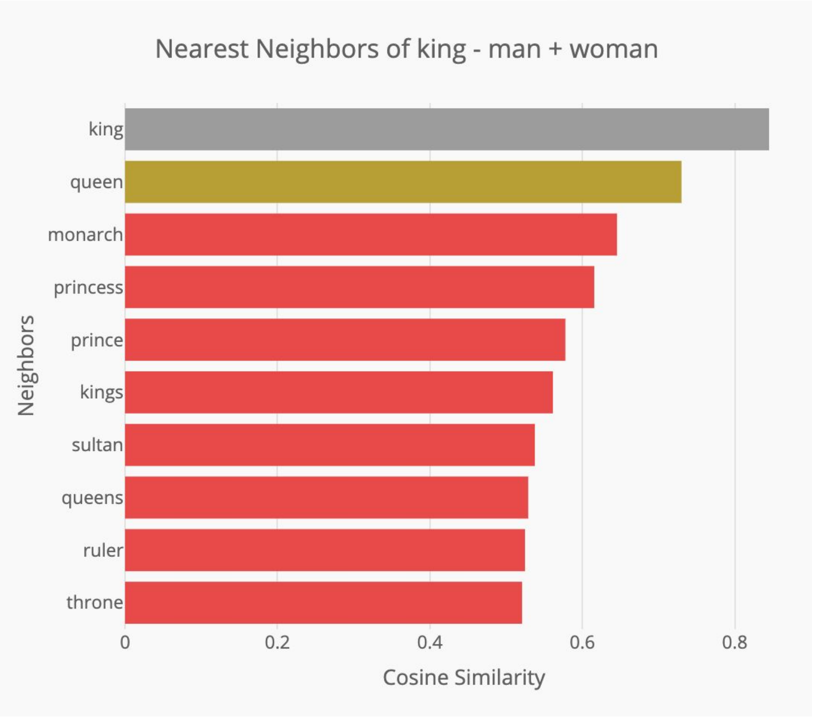
spain - madrid + berlin => Germany

california - state + country => USA

man - king + woman => queen

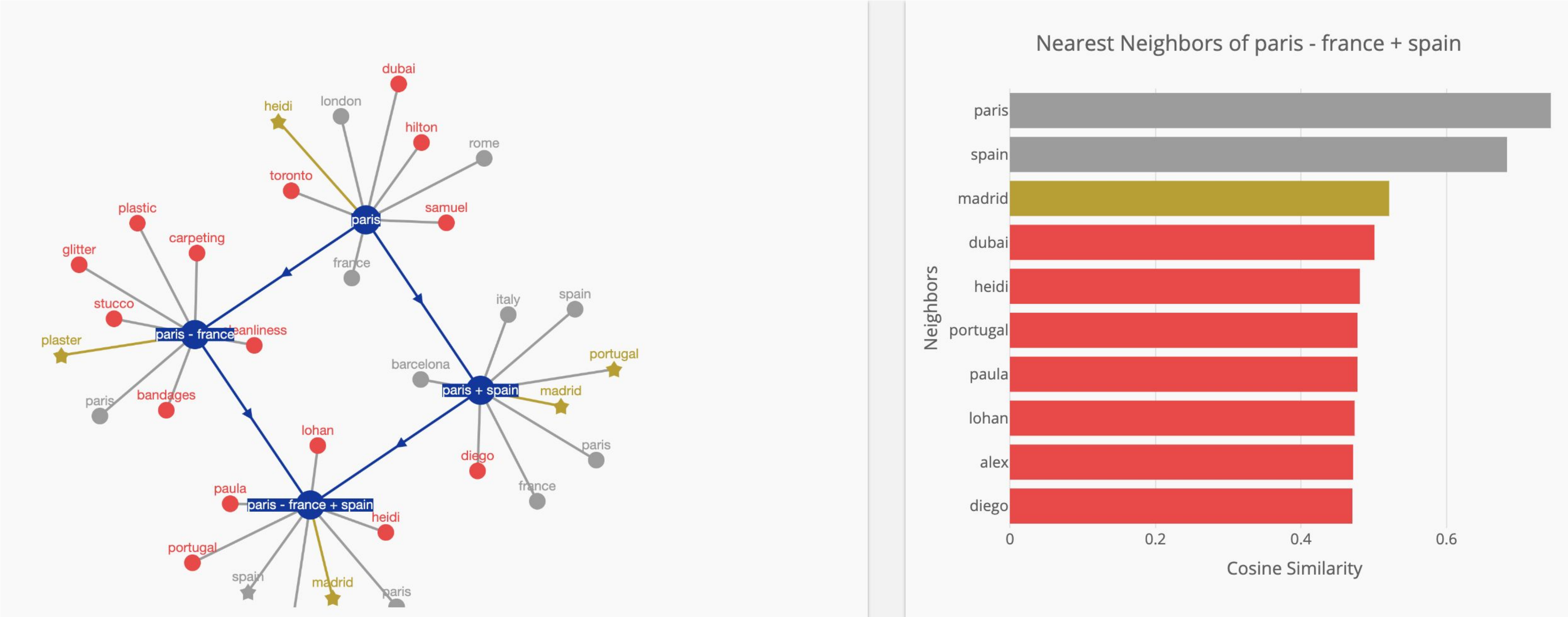
Word vectors operations





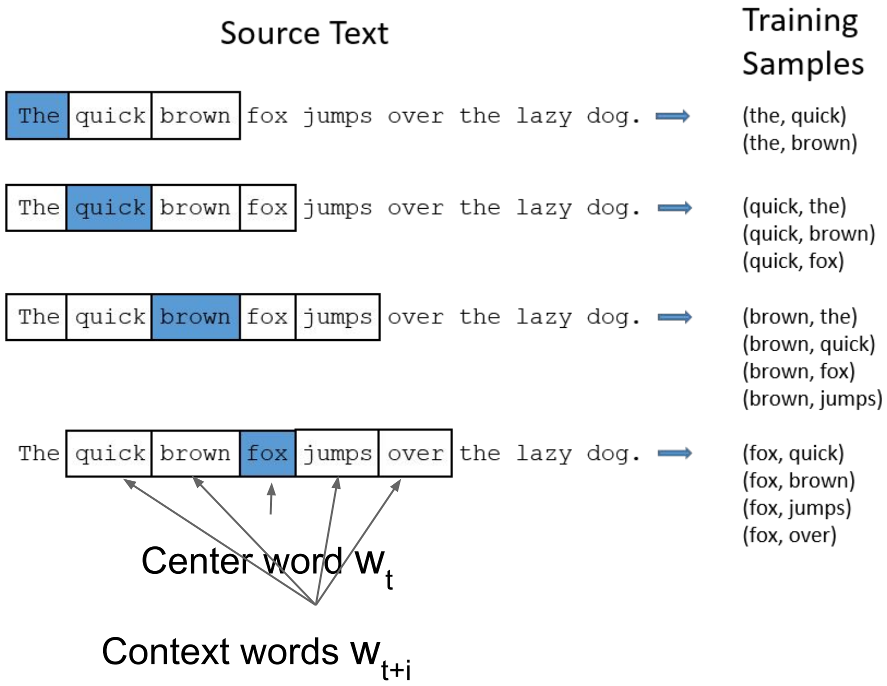
https://dash-gallery.plotly.host/dash-word-arithmetic/

Word vectors operations



https://dash-gallery.plotly.host/dash-word-arithmetic/

Word2vec



Source Text

Training Samples

The quick brown fox jumps over the lazy dog. → (the, quick) (the, brown)

The quick brown fox jumps over the lazy dog. → (quick, the) (quick, brown) (quick, fox)

The quick brown fox jumps over the lazy dog. → (brown, the) (brown, quick) (brown, fox) (brown, jumps)

The quick brown fox jumps over the lazy dog. → (fox, quick) (fox, brown) (fox, jumps) (fox, over)

Center word w_t

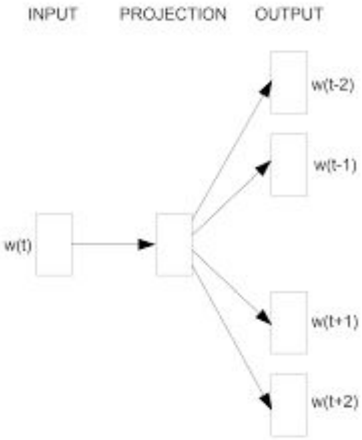
Context words w_{t+i}

- Make windows with size m, with center words (w_t) and context words (w_{t+i})
- And model: $P\left(w_{t+j} \mid w_t\right)$

$$J(\theta)=\prod_{t=1}^{T} \prod_{\substack{-m \leq j \leq m \\ j \neq 0}} P\left(w_{t+j} \mid w_t ; \theta\right)$$

- How to model:
  - vector central: v_c
  - vector contexts: u_o

Word2vec: skip-gram



Skip-gram

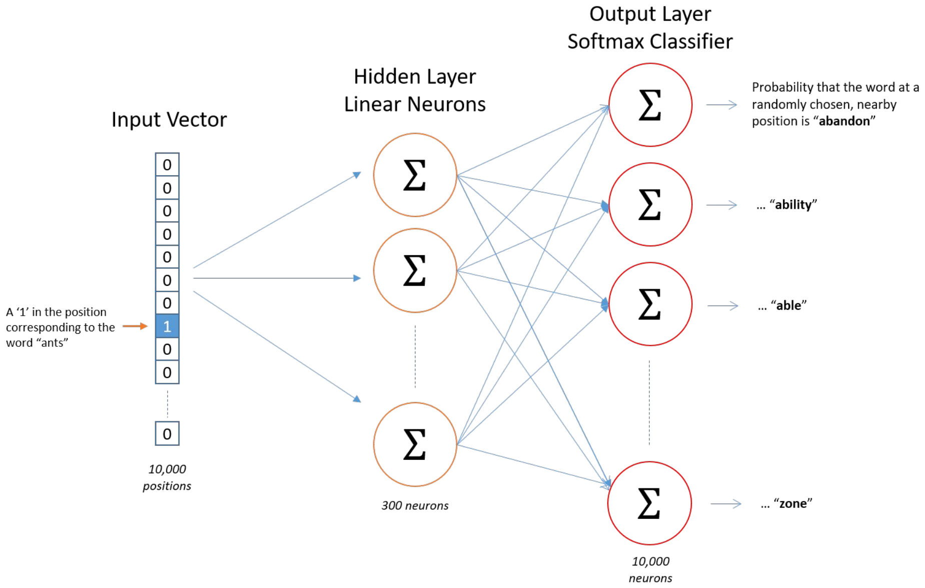
- The distributed representation of the **central** word is used to predict the **context**.

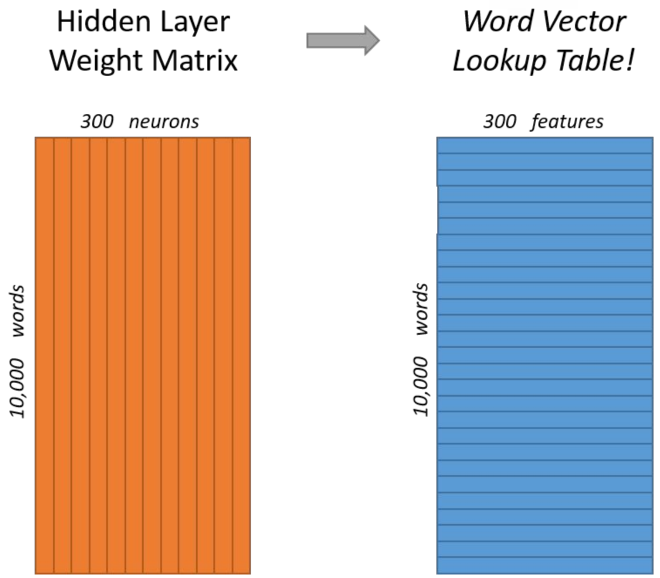
$$P\left(w_{t+j} \mid w_t\right)$$

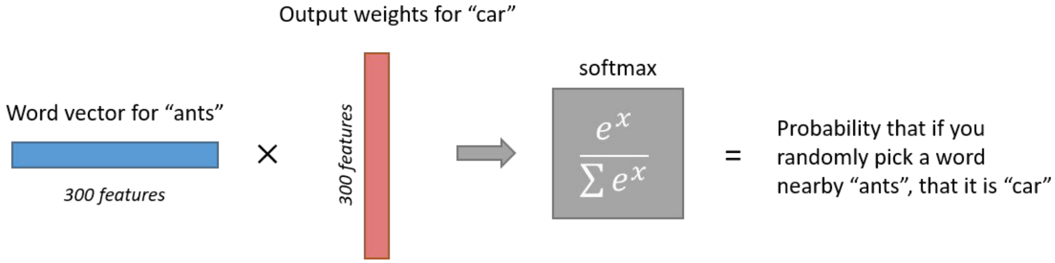
- For central word c and context word o (softmax):

$$P(o \mid c)=\frac{\exp \left(u_o^T v_c\right)}{\sum_{w \in V} \exp \left(u_w^T v_c\right)}$$

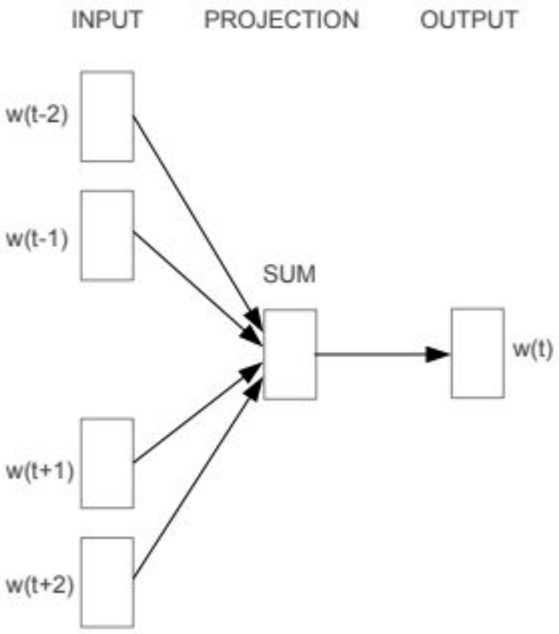
Word2vec: skip-gram


Input Vector
0
0
0
0
0
0
0
1
0
0
0
10,000 positions
A '1' in the position corresponding to the word "ants"
Hidden Layer Linear Neurons
Σ
Σ
Σ
300 neurons
Output Layer Softmax Classifier
Σ
Σ
Σ
Σ
10,000 neurons
Probability that the word at a randomly chosen, nearby position is "abandon"
… "ability"
… "able"
… "zone"





Word2vec: CBOW



CBOW

In the CBOW model, the distributed representations of context (or surrounding words) are combined to predict the word in the middle.

GloVe: Global Vectors

- Incorporates global statistics.
- Co-occurrence matrix: X_ij Number of times word i, appears with word j
  - Window based
  - Document based
- GloVe outperforms word2vec when the corpus is small or where insufficient data may be available to capture local context dependencies.

$$J = \sum_{i=1,j=1}^{V} f(X_{ij}) \left(\mathbf{u}_i^{\top} \mathbf{v}_j - \log\left(X_{ij}\right)\right)^2$$

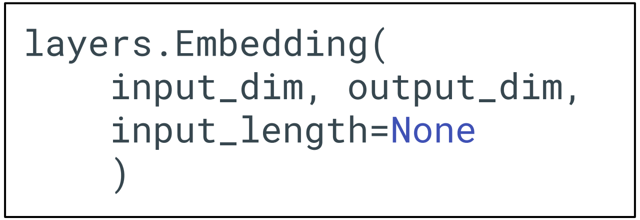
FastText: Subword Embeddings

- FastText is an extension to Word2Vec proposed by Facebook in 2016.
- Word2Vec and GloVe embedding can't encode unknown words.
- FastText breaks words into several n-grams (sub-words). Example: *apple* is *app*, *ppl*, and *ple*
- Word embedding vector for *apple* will be the sum of all these n-grams

Amígdala - Amigdalitis
Casa - Casoplón
Google - Googlear

Embeddings in a Neural Network

- You can use embeddings for encoding any categorical entity. Words, products in a recommender system, clients, films… (limitations of One Hot Encoding)
- You can use them for visualization of concepts and relations between categories.
- **Learning Embeddings:** weights of the network, look-up table.



```
layers.Embedding(
    input_dim, output_dim,
    input_length=None
    )
```

- input_dim: This is the size of the vocabulary.
- output_dim: Vector space dimension
- input_length: length of input sequences

Exercise

ntic
master
revolucionamos la comunicación

UNIVERSIDAD
COMPLUTENSE
DE MADRID