Introduction to natural language processing for health and biological questions

Nona Naderi
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique

TF-IDF representation of documents

Weigh the importance of each word with the document frequency.

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t.$$

Assignment 1

Use the following dataset https://github.com/sb895/Hallmarks-of-Cancer/tree/master, describe it in terms of annotations and number of documents, tokens (using scispacy), and annotations. Represent the articles using TF-IDF.

Pointwise Mutual Information (PMI)

Instead of absolute co-occurrence statistics, use probability of co-occurrences

$$\mathrm{PMI}(w_1 w_2) = \log \frac{p(w_1 w_2)}{p(w_1)p(w_2)}$$

Vector-based semantics

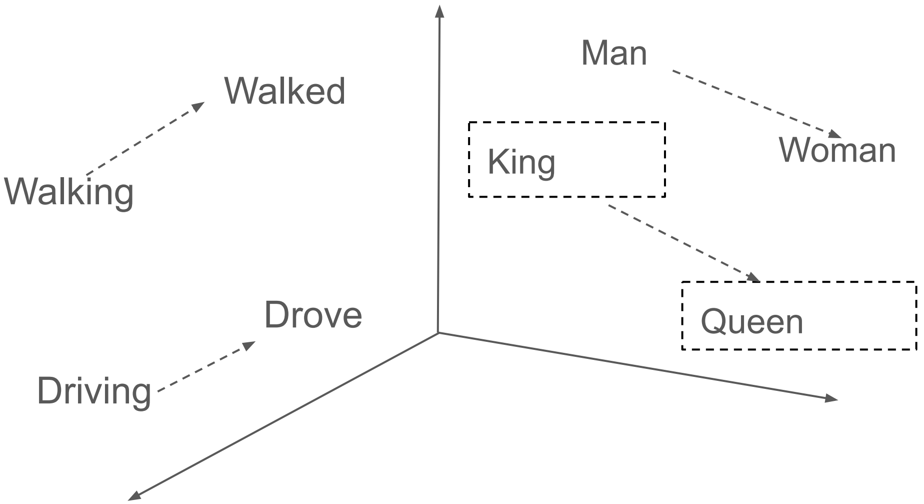
You can get a lot of information by representing a word using it context.

“Explore **bank** accounts, loans, mortgages, investing, credit cards & **banking** services”

“Measurements of **bank** erosion on rivers in Devon over a 2½ year period produced mean rates ranging from 0.08 to 1.18 metres per year and a maximum rate of 2.58 metres per year.”

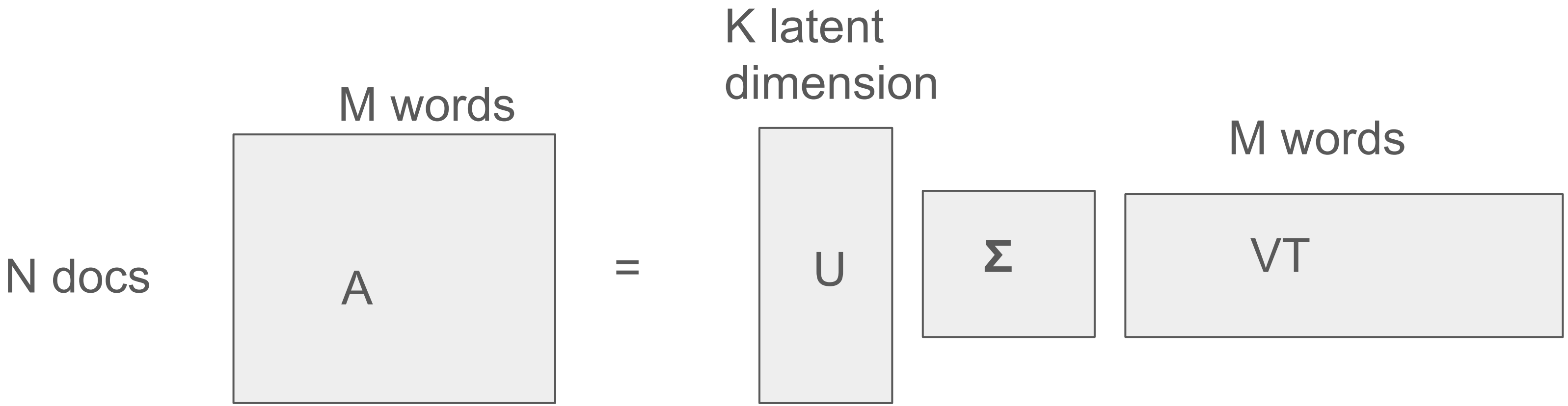
Distributed, distributional representations

With these representations, syntactic and semantic patterns are captured.



Count vs. Predict

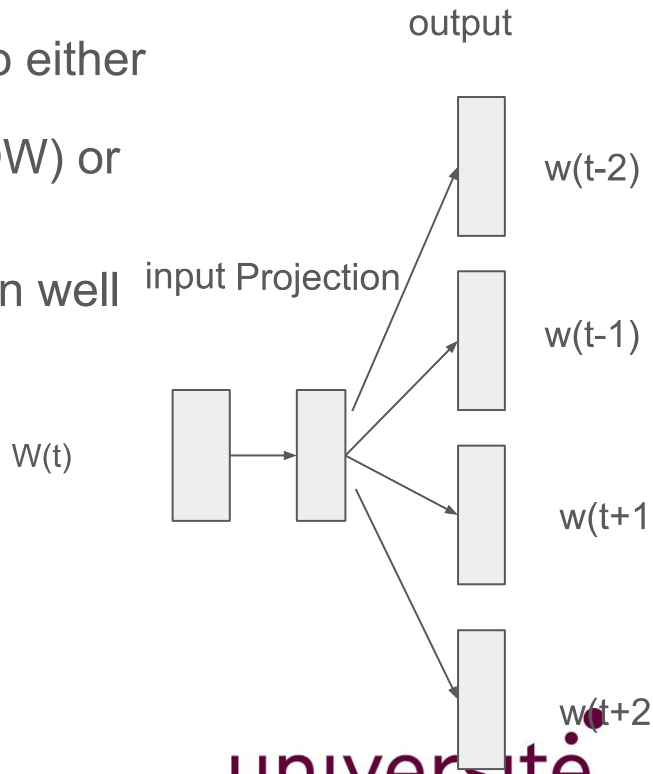
Latent Semantic Analysis (LSA): Term-document co-occurrence matrix with dimensionality reduction using Singular Value Decomposition (SVD)



Word2vec

Word2vec CBOW/SkipGram: Train word vectors to try to either

- Predict a word given its bag-of-words context (CBOW) or
- Predict a context word from the center word
- Update word vectors until they can do this prediction well



Exercise

Use pre-trained word2vec model (google-news-300) to find the most similar terms to dog. (you can use Gensim package)

Language model

A language model is a probability distribution over words or word sequences.

What is the probability of “ I go to Paris-Saclay University”?

What is the probability of “To go I University Paris-Saclay”?

$P(w) = P(w_1 w_2 w_3 w_4 \ldots w_k)$

$P(w_k \mid w_2 w_3 \ldots w_{k-1})$

N-grams models

Unigram model: $P(w_1)P(w_2) \dots P(w_n)$

Bigram model: $P(w_1)P(w_2|w_1)P(w_3|w_2)...P(w_n|w_{n-1})$

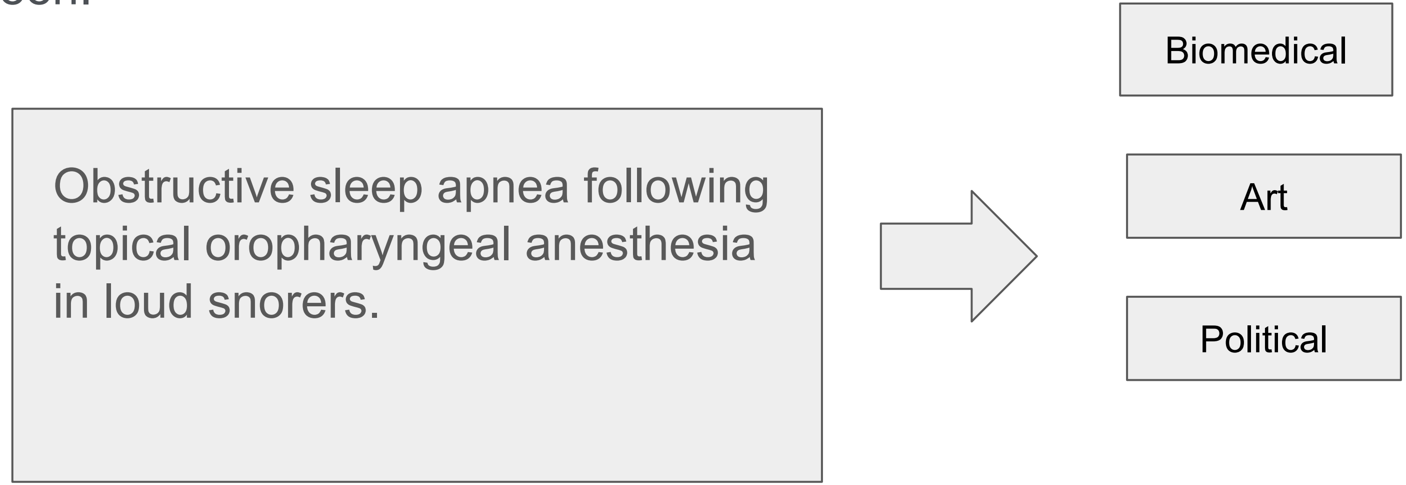
Trigram model: $P(w_1)P(w_2|w_1)P(w_3|w_2,w_1).....P(w_n|w_{n-1}w_{n-2})$

Markov chain

- A Markov model of *order 0* predicts that each letter in the alphabet occurs with a fixed probability.
- We can fit a Markov model of order 0 to a specific piece of text by counting the number of occurrences of each letter in that text, and using these counts as probabilities.
- Example: “baggage” p(a) = 2/7

Text classification

- A classifier is a function that maps inputs to a predefined sets of class labels.
- Our classifier needs to be able to classify items that our model has never seen.



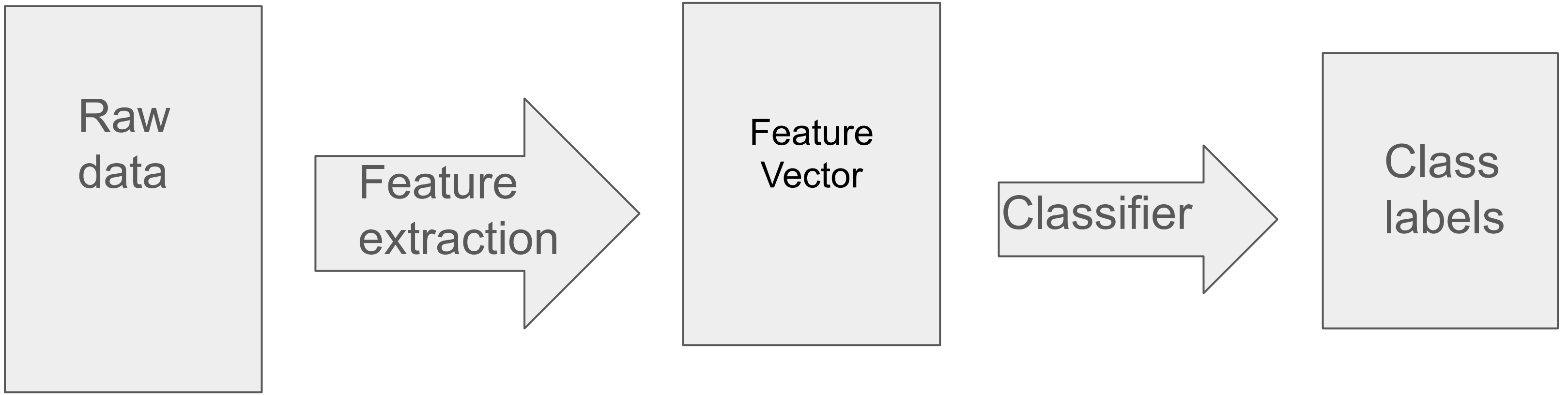
Obstructive sleep apnea following topical oropharyngeal anesthesia in loud snorers.

Biomedical

Art

Political

Text classification



Raw data

Feature extraction

Feature Vector

Classifier

Class labels

Evaluation

Precision: What percentage of items that were assigned label X do actually have label X in the test data?
Recall: What percentage of items that have label X in the test data were assigned label X by the system?
F-Measure: harmonic mean of precision and recall
$F = (2 \cdot P \cdot R) / (P + R)$