POLIMI GRADUATE SCHOOL OF MANAGEMENT

INTRODUCTION TO NATURAL LANGUAGE PROCESSING

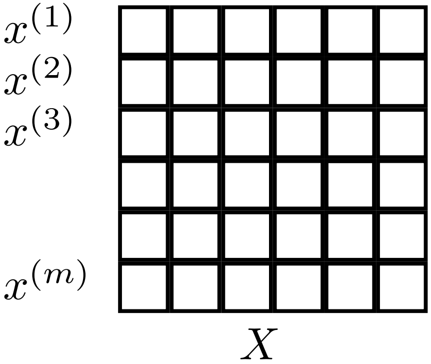
Andrea Mor - andrea.mor@polimi.it

CONTENTS

- Introduction
- Data Preparation
- Sentiment Analysis
- Topic Modeling

NATURAL LANGUAGE PROCESSING

- From structured data to unstructured data

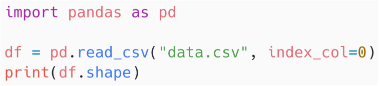


$x^{(1)}$: Walking across the sitting room, I turn the television off.

$x^{(2)}$: @UnitedAirlines second flight in a month #cancelled thanks a lot #worstairlineever

$x^{(3)}$: $E = mc^2$

$x^{(4)}$:



```
import pandas as pd

df = pd.read_csv("data.csv", index_col=0)
print(df.shape)
```

SOME IMPORTANT QUESTIONS

- We would like to exploit “all” that we know on structured data
- How can we represent text in a structured way?
- Which are the limits/problems of a representation?

THE ELEMENTARY UNIT

The elementary unit of a digital image is the pixel, what about text?

→ words? letters? tokens?

- The same words can have different meanings
- The same meaning can be expressed by different words

CONTEXT AND SEQUENCES

- The meaning of a word depends on its context
- Context is not always defined by proximity

This summer I went to Italy. I had a lovely time there, engaging in both recreational and cultural ***<blank#1>****, meeting old friends and making new ones [...] but most of all I had my food,* ***<blank#2>****.*

CONTEXT AND SEQUENCES

- The meaning of a word depends on its context
- Context is not always defined by proximity

  *This summer I went to Italy. I had a lovely time there, engaging in both recreational and cultural **<blank#1>**, meeting old friends and making new ones [...] but most of all I had my food, **<blank#2>**.*

- Order plays an important role

  ***Only he** can play that instrument*
  ***He only** can play that instrument*

TEXT CLEANING

- Convert to lower case
- Remove punctuation
- Remove numerical values
- Typos
- Remove special characters ([?@)
- Remove stop words (the, it, etc)
- Remove special description words ([chorus], [fade], [applause])
- Tokenize text (O'Neill → [o] [neill], [o'neill]?; aren't → [arent], [are][nt]?)
- Create bi-grams or tri-grams ([United Kingdom] vs [United][Kingdom])
- Normalization:
  - Stemming (car, cars, car's, cars' → car;)
  - Lemmatization (am, are, is → be)

TEXT REPRESENTATION

1. Corpus: a collection of text
2. Document-Term Matrix: word counts in matrix format
3. TF-IDF: Term Frequency - Inverse Document Frequency

$$\text{TF-IDF} = f_{t,d} \times \text{idf}(t, D)$$

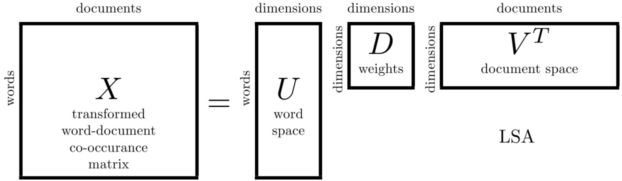
where

- $f_{t,d}$: the number of times that term t occurs in document d.
- $\text{idf}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}| + 1}$: log of the inverse of the number of documents containing term t.

SENTIMENT ANALYSIS

- TextBlob Module: Linguistic labeled the sentiment of words. https://github.com/sloria/TextBlob/blob/eb08c120d364e908646731d60b4e4c6c1712ff63/textblob/en/en-sentiment.xml
- Sentiment Labels: Each word is labeled in terms of
  - Polarity: negative(-1) or positive(+1)
  - Subjectivity: subjective(0) or fact(+1)
- Sentiment of words can vary based on where it is in a sentence.
  - Negation multiplies the polarity by -0.5

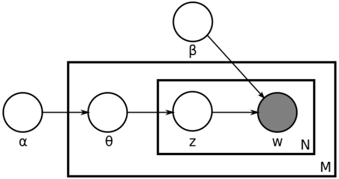
TOPIC MODELING - LSA



Latent Semantic Analysis (LSA)

- **Singular Value Decomposition (SVD)** of the Document-Term Matrix

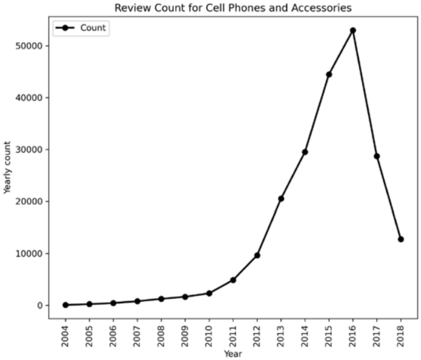
TOPIC MODELING - LDA

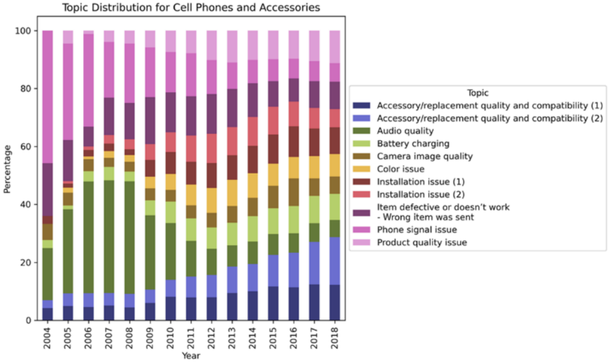


Latent Dirichlet Allocation (LDA)

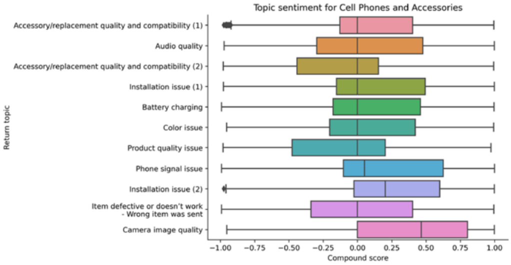
- **Documents (M) are probabilistic distribution over topics**: let say that a document is p_i% of topic i.
- **Topics are probabilistic distribution over words**: given a topic chosen according to the distribution of the document, we generate a word according to the topic distribution
- Random initialisation: assign each word to a random topic
- Update each word by considering
  - proportion of words in the document of topic
  - proportion of topics in all documents for the word
- Repeat until stopping condition

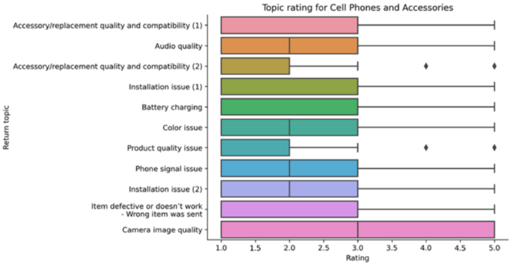
METADATA: TOPICS IN TIME





TOPIC + SENTIMENT AND METADATA


Topic sentiment for Cell Phones and Accessories


Topic rating for Cell Phones and Accessories

THANK YOU