slidenumbers: true
footer: A. Ferrara. **Language models**. *Part 1: introduction*. email, corse website, slack, github

Information Retrieval

[fit]Language models

Part 1: Introduction. Prof. Alfio Ferrara

Master Degree in Computer Science

Master Degree in Data Science and Economics

Introduction

A **language model** in essentially a probability distribution over a sequence of words

$p(w_1, w_2, \ldots, w_n)$

which can be used for a surprisingly high number of tasks, including *document search*, *document classification*, *text summarization*, *text generation*, *machine translation*, and many others

Note: Instead of estimating the probability distribution of words, we can work at a finer granularity on the distribution of substrings of fixed length in words (e.g., characters, 2-chars blocks)

Example 1

A LM may be used to guess the next word in a sequence

$p(w_n|w_1, w_2, \ldots, w_{n-1})$

> Yesterday $\rightarrow$ you $\rightarrow$ studied,
> what $\rightarrow$ are $\rightarrow$ you $\rightarrow$ going $\rightarrow$ to $\rightarrow$ do $\rightarrow$ ... ?
> $\rightarrow$ today

Example 2

Or to guess the author (or any other categorical attribute) of as text

$p(author \mid w_1, w_2, \ldots, w_n)$

> "Twenty years from now you will be more disappointed
> by the things that you didn't do than by the ones you did do"
> $\rightarrow$ Mark Twain

Example 3

Or to select the correct translation for a sentence

$p(english \mid option1)\ vs\ p(english \mid option2)$

> "ci sono molti esempi"
> $\rightarrow$ *"there are many examples"*
> $\rightarrow$ ~~"are there many examples"~~

Types of Language Models

Two main types of LM:

1. **Statistical Language Models**: Estimate the probability distribution of words by enforcing statistical techniques such as n-grams *maximum likelihood estimation (MLE)* or *Hidden Markov Models (HMM)*
2. **Neural Language Models**: Popularized by [1], each word is associated with an embedding vector of fixed size and a Neural Network is used to estimate the next word given a sequence of k preceeding words

---

[.autoscale: true]

Outline of the thematic study

1. **Part 1: Introduction**
    1. A case study to motivate the need of Language Models and to serve as a running example
2. **Part 2: Bases and evaluation**
    1. Some basics about estimating probabilities and n-grams
    2. Strategies for evaluating Language Models
3. **Part 3: Statistical Language Models**
    1. Statistical Language Models and their limitations
    2. Topic modeling
4. **Part 4: Word embedding**

    1. Word embedding models
    2. Usage of word embedding
5. **Part 5: Neural Language Models**
    1. Introduction to Neural Language models
6. **Part 6: Applications**
    1. Language Models for text classification and generation

---

[fit] Acting Lessons: The Cornell Movie-Dialogs Corpus

[fit] Case Study

The **The Cornell Movie-Dialogs Corpus** contains a large metadata-rich collection of fictional conversations extracted from raw movie scripts [2] .

Our aim is to model the language of fictional characters and the language of a whole movie, with the task of measuring similarities, search, classification, and generation of fictional text.

MONTY
They just left him here to die. They threw him out the window and kept driving.

A ship's horn sounds from the Hudson.

KOSTYA
Come, my friend, it is cold. Come, people wait for us.

MONTY
They're used to waiting.

Monty squats near the dog and inspects him. From this angle it is clear that the pit bull has been badly abused. One ear has been chewed to mince; his hide is scored with cigarette burns; flies crawl in his bloodied fur.

MONTY (CONT'D)
I think maybe his hip—

The dog pounces, jaws snapping,; lunging for Monty's face.

Monty stumbles backwards. The dog, too badly injured to continue the attack, remains in his crouch, growling.

Monty sits on the pavement, shaking his head.

MONTY (CONT'D)
Christ.
(beat)
He's got some bite left.

KOSTYA
I think he does not want to play with you. Come, you want police to pull over? You want police looking through your car?

MONTY
Look what they did to him. Used him for a fucking ashtray.

Monty stands and dusts his palms on the seat of his pants.

Main figures

220,579 conversational exchanges between 10,292 pairs of movie characters

involves 9,035 characters from 617 movies

in total 304,713 utterances

movie metadata included:

- genres
- release year
- IMDB rating

- number of IMDB votes
- IMDB rating

character metadata included:

- gender (for 3,774 characters)
- position on movie credits (3,321 characters)

---

movie_titles_metadata.txt

contains information about each movie title

- fields:
    - movieID,
    - movie title,
    - movie year,
    - IMDB rating,
    - no. IMDB votes,
    - genres in the format `['genre1','genre2',…,'genreN']`

---

movie_characters_metadata.txt

contains information about each movie character

- fields:
    - characterID
    - character name
    - movieID
    - movie title
    - gender ("?" for unlabeled cases)
    - position in credits ("?" for unlabeled cases)

---

movie_lines.txt

contains the actual text of each utterance

- fields:
    - lineID
    - characterID (who uttered this phrase)
    - movieID
    - character name
    - text of the utterance

---

movie_conversations.txt

the structure of the conversations

- fields
  - characterID of the first character involved in the conversation
  - characterID of the second character involved in the conversation
  - movieID of the movie in which the conversation occurred
  - list of the utterances that make the conversation, in chronological order: `['lineID1','lineID2',…,'lineIDN']`
    has to be matched with movie_lines.txt to reconstruct the actual content

raw_script_urls.txt

the urls from which the raw sources were retrieved

Dataset acquisition

The dataset has been uploaded in `MongoDb` using the script available here.

The database `movie-dialog` is composed of 4 collections:

- `movies`: storing movie metadata
- `characters`: storing metadata about characters in movies
- `line`: storing lines played by characters in movies
- `conversations`: sequences of lines that constitute a conversation between two characters in a movie

---

Lines data


```
{
"_id" : ObjectId("5e9b0c10af9b291d85c99978"),
  "character" : {
    "name" : "BIANCA",
    "movie" : {
      "title" : "10 things i hate about you",
      "year" : 1999,
      "rating" : 6.9,
      "votes" : 62847,
      "genres" : [
        "comedy",
        "romance"
      ],
      "id" : "m0"
    },
    "gender" : "f",
```

```
    "pos" : 4,
    "id" : "u0"
  },
  "text" : "They do not!",
  "id" : "L1045"
}
```

Conversation data

[.code-highlight: 9, 11, 13, 15]

```
{
  "_id" : ObjectId("5e9b1168af9b291d85cf845a"),
  "character_a" : {"name" : "BIANCA", "gender" : "f", "pos" : 4, "id" : "u0"},
  "character_b" : {"name" : "CAMERON", "gender" : "m", "pos" : 3, "id" : "u2"},
  "movie" : {"title" : "10 things i hate about you", "year" : 1999, "rating" : 6.9,
    "votes" : 62847, "genres" : ["comedy", "romance"], "id" : "m0"},
  "lines" : [
    {"line" : "L194", "character" : "u0", "gender" : "f",
     "text" : "Can we make this quick?  Roxanne Korrine and Andrew Barrett are having an incredibly horrendous public break- up on the quad.  Again."},
    {"line" : "L195", "character" : "u2", "gender" : "m",
     "text" : "Well, I thought we'd start with pronunciation, if that's okay with you."},
    {"line" : "L196", "character" : "u0", "gender" : "f",
     "text" : "Not the hacking and gagging and spitting part.  Please."},
    {"line" : "L197", "character" : "u2", "gender" : "m",
     "text" : "Okay... then how 'bout we try out some French cuisine.  Saturday?  Night?"}
  ],
  "len" : 4
}
```

Main figures (I)

Movies

genre	movies	avg rating	avg votes	min year	max year
crime	147	7.00476	62577	1932	2009
horror	99	6.06263	26833.7	1931	2010
fantasy	78	6.59615	55093.1	1932	2009
adventure	116	6.87155	72374.5	1927	2007
sci-fi	120	6.66833	66578.6	1927	2009
mystery	102	6.96961	63826	1933	2009
romance	132	7.05076	41146	1927	2005
thriller	269	6.71413	57825	1933	2010
action	168	6.55833	61224.8	1949	2009
drama	320	7.23656	52887.8	1927	2010
comedy	162	6.7284	35270.9	1931	2007
others	159	6.97273	39685.2	1927	2010
total	**1872**	**6.78616**	**52943.6**	**1927**	**2010**

Main Figures (II)

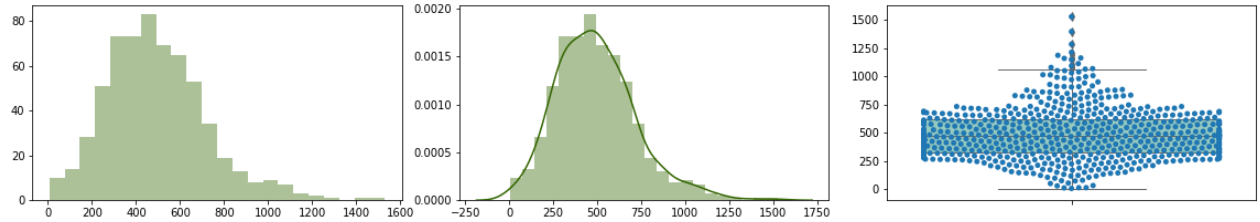
Characters and lines

	Number	Avg movie	Male	Female	Unknown
characters	9035	14.64	2049	966	6020

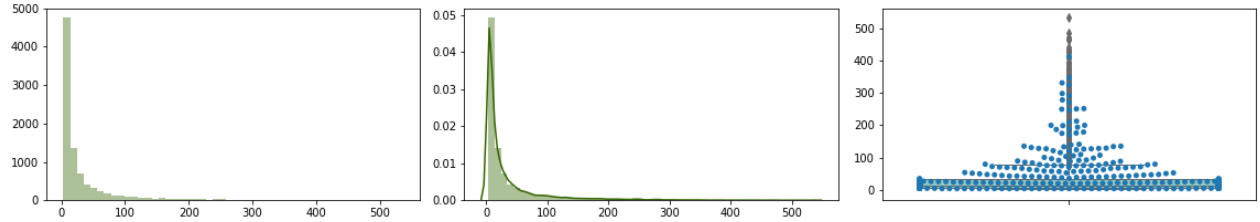
	Number	Avg movie	Avg character
lines	304 713	162.8	33.7

Main Figures (III)

Distributions of lines per movie

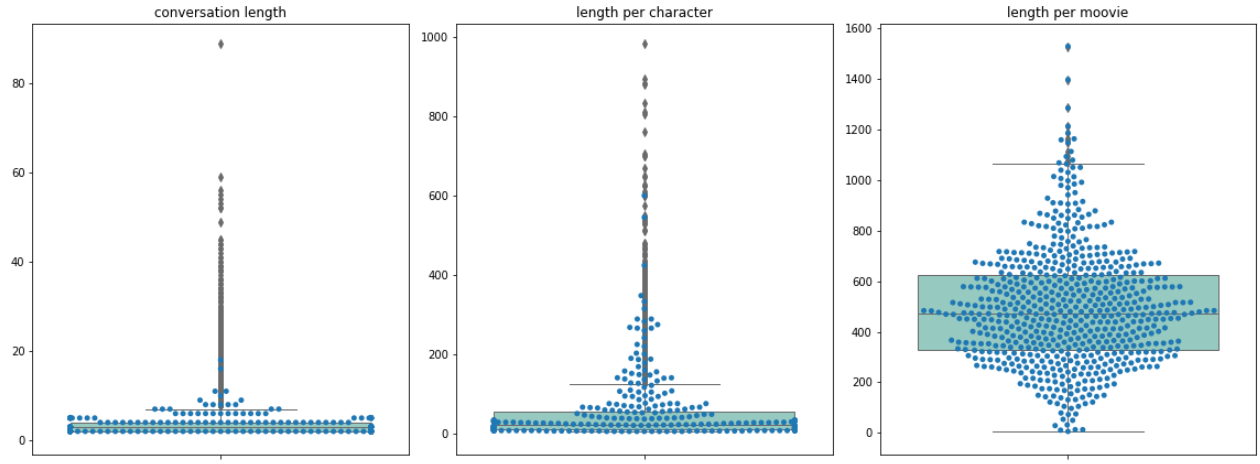


Distribution of lines per character



Main Figures (IV)

Lenght of conversations



Some considerations

- In general, there are few lines per character. This makes it difficult to statistically modeling the language of single characters.
- There are several levels of possible aggregation of lines (e.g., by movie, by gender, by year, by genre, and combinations of those). This helps in building larger corpora suitable for statistical modeling.
- Conversations may be useful to build models for predicting the dialogue options.
- Gender and year are interesting options for working on multi class classification, while movie genre can be used for multi-label classification and for topic modeling.

1. Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, *3*(Feb), 1137-1155. ↩

2. Danescu-Niculescu-Mizil, C., & Lee, L. (2011, June). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In Proceedings of the 2nd workshop on cognitive modeling and computational linguistics (pp. 76-87). Association

for Computational Linguistics. ↩