Similarities between researchers

Luis Galdo Seara

Index

Datasets

- NIPS
  - The Conference and Workshop on Neural Information Processing Systems is a machine learning and computational neuroscience conference
  - 6560 papers and 8653 authors
- ACL
  - The ACL Anthology hosts papers on the study of computational linguistics and natural language processing
  - 22460 papers and 14616 authors

Objective

- One function for each dataset
  - Given the name of an author from one dataset return related authors from the other dataset based on the main topic of each author

Procedure: Loading NIPS dataset

- Originally a CSV
- Transformed into a JSON to avoid problems with Windows
- Loaded using pandas
- Main information of the data frame
  - Id, title, year and text

Procedure: Loading ACL dataset

- Folder containing a .txt file for each paper
- Loaded into a list of lists
  - Each list has an id and the text corresponding to that id

Procedure: Cleaning using NLTK

- Split into words and convert into lower case
- Remove:
  - Punctuation
  - Non alphabetic tokens
  - Stop words
- Stemming
- Removing words of less than 3 characters

Procedure: TFIDF

- Merging clean papers from both datasets
- TFIDF
  - Max_df = 0.95
  - Min_df = 0.2
  - Max_features = 100

Procedure: NMF Topic Modeling

- Previous experience
- Why NMF in this case?
  - More clear
  - More intuitive results

Procedure: NMF Topic Modeling

- 8 different topics
- Why?
  - Easier to assign names and to see the differences
  - Both datasets are related to Machine Learning
- Each author is assigned to a topic based on the topics of the papers he published

Procedure: Topics

- Information Extraction
- Stochastic Methods
- Parsing Techniques
- Probabilistic Methods
- Reinforcement Learning
- Translation Techniques
- Words Segmentation
- Training Neural Networks

Results

- Two files (one related to each dataset) containing the ids of each author and their main topic
- Using the files that relate the ids of each author to their names, the objective function is built

Thank you!

- Any Questions?
- GitHub