Identifying document topics using the Wikipedia category network

Web Information Retrieval project

Artuso Fiorella 1602113
Migliori Andrea 1607771

SAPIENZA
UNIVERSITÀ DI ROMA

Introduction

- The goal of such an experiment is to show that it is possible to identify quite well the Wikipedia categories most characteristic of a document even with a simple algorithm that exploits only titles, redirections and categories of Wikipedia articles.

- In fact, each Wikipedia article consists of:
  - **title**
  - set of **categories**
  - set of **redirections**

- Our entire work is organized into three main steps:
  1. creation of the dataset
  2. implementation of the algorithm
  3. Validation of the results

Experimental setting

In our work we made two main choices that differ from the paper:

- **Reduced size of the dataset**
  - the current Wikipedia snapshot is much greater than the one used in the paper.
  - it is infeasible to process all the currently available articles (> 5 years)
  - we worked on a restricted dataset of **33.500 articles**

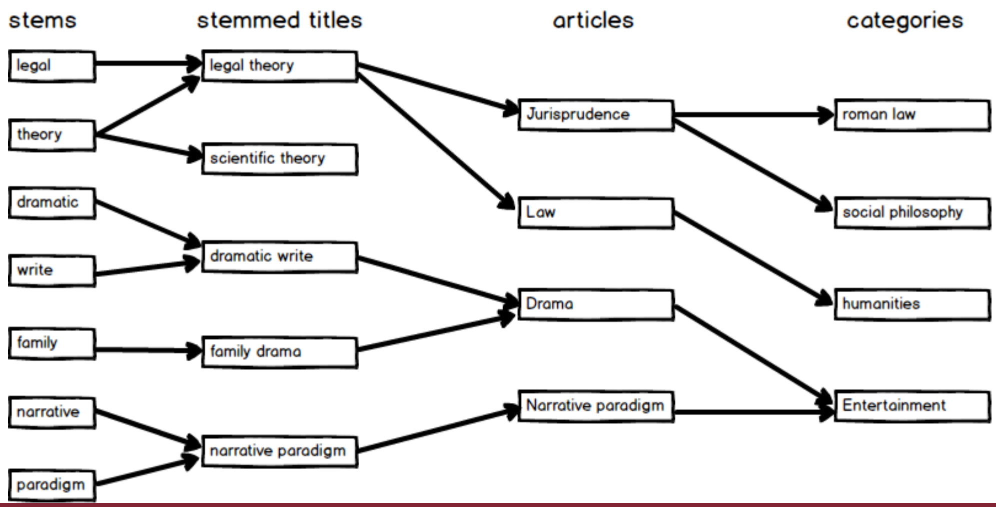
- **Reduced number of categories**
  - risk of selecting a high number of very heterogeneous categories – the algorithm would then poorly identify document topics.
  - solution:
    - taking one category among the list of Wikipedia's major topic classifications (such as arts, business, sport, religion, etc…)
    - selecting at random either the previously chosen category itself or one the its subcategories up to 3 level of indirection
    - randomly picking one Wikipedia article having the selected category among its categories set.

Creation of the dataset

- redirections and categories are retrieved from DBpedia (represented through RDF) by using SPARQL queries

- in order to add an article to the corpus, each article undergoes the following steps:
  - extract redirections and categories
  - perform stop word removal and stemming on article title and redirections.
  - remove categories corresponding to Wikipedia administration and maintenance
  - remove categories containing less than 5 articles
  - merge stub categories with regular ones

- the resulting dataset is structured as follows:



Identifying document topics

We identify topics of an input document through a series of steps:

- ***step 1*** – stop words removal and stemming on the input document

  NB: we are going to consider only words present both in the document and in the dataset

- ***step 2*** – assign a weight to each word*:* $R_w = tf_w \times \log \frac{N}{cf_w}$

- ***step 3*** – collect stemmed titles supported by words present in the document and weight them:

$$R_t = \sum_{w \to t} R_w \times \frac{1}{t_w} \times \frac{1}{a_t} \times \frac{S_t}{L_t}$$

- ***step 4*** – collect articles pointed to by the titles found in the previous step and weight them:

$$R_a = \max_{t \to a} R_t$$

- ***In step 5*** **–** collect categories associated to the articles above and weight each of them:

$$R_c = \sum_{a \to c} R_a$$

First improvement: since a category can have an high weight due to the presence of many titles pointing to articles in that category, smooth this by modifying the formula above:

$$R_c = \frac{v_c}{d_c} \times \sum_{a \to c} R_a$$

Second improvement: since a supporting word may appear in the vocabulary of many categories it would contribute to their weights in exactly the same way but it would be better for such supporting word to contribute differently to each of the categories whose vocabularies contain it.

$$R'_c = R_c \times \frac{\sum_{w \in B_c} d_w}{|B_c|} \qquad d'_w = \frac{d_w}{2}, w \in B_c$$

- ***In step 6*** – select the top 20 categories

Validation of the results (naïve test)

- It aims to test whether our algorithm is good in predicting the **central topic** of a given document by observing the top 20 categories assigned to it.
- Such input documents are selected from the 20 Newsgroups dataset

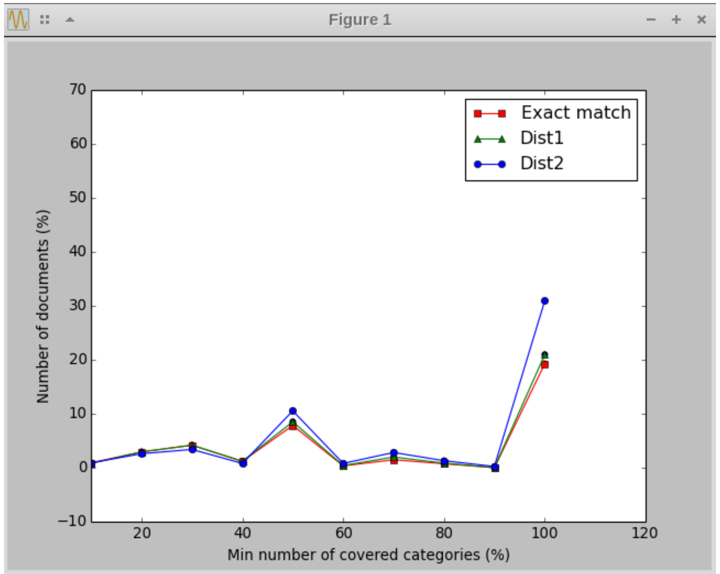
PRE-OPT	OPT1	OPT2
Religion: 37.05451373682791	Atheism: 2.6921802698543633	Atheism: 2.6921802698543633
Philosophy: 35.12936723774073	Skepticism: 2.3905004638483325	Skepticism: 1.6932711618925689
Language: 35.01543176590475	Philosophy: 1.7717419998164892	Philosophy: 1.6190056205219643
Life: 32.789054970281136	20th Century Fox films: 1.7617717507354547	20th Century Fox films: 1.4814898813002688
Entertainment: 25.216036981868772	Criticism of religion: 1.627509063266141	Religion: 0.8520139465581635
Mathematics: 24.606303401321014	Humanism: 1.590082513244776	Humanism: 0.7232666987328669
Culture: 23.998614592309835	Religion in science fiction: 1.5543366334022737	Language: 0.6744504458044537
Science: 23.21907916211353	Søren Kierkegaard: 1.4654113367217068	Criticism of religion: 0.6200034526728156
Business: 20.718455778628066	2010s thriller films: 1.4614737164603169	Søren Kierkegaard: 0.54952925127064
Living people: 19.851070046393627	Religion: 1.4519527609985272	2010s science fiction films: 0.4784983196730488
Law: 18.706224847939733	Agnosticism: 1.387523534425303	Religion in science fiction: 0.47185219228283304
Politics: 17.48351855466089	2010s science fiction films: 1.319995364615307	Life: 0.46452479944021596
Sports: 16.723041425670747	Films about religion: 1.316822925946283	Culture: 0.3984840607316937
Technology: 16.292607938792003	Philosophical movements: 1.3142527675337132	2010s thriller films: 0.3653684291150792
Society: 16.02152394280103	Space adventure films: 1.2934186919975765	Space adventure films: 0.33851192329624075
History: 14.613003560780115	Irreligion: 1.2904276650775248	Mathematics: 0.27537697449825915
Concepts: 14.09205077445818	Language: 1.2679902191431995	American films: 0.26457150155370435
Reference: 13.422759302194454	Secularism: 1.1910682016011611	Philosophical movements: 0.2592568935955176
Nature: 12.901784001880241	Life: 1.1815875664966176	Science: 0.23822797895210296
Education: 12.892386207195324	Culture: 1.1752969797866306	Collective rights: 0.229677263171798

This table shows the results of our algorithm on the document "49960" which is about atheism

Validation of the results (real test)

- In order to measure how well our algorithm can predict the **original categories**, we run it on the body of 1775 Wikipedia articles not contained in the dataset and randomly selected from Wikipedia in exactly the same way as we did for the creation of the dataset.

NB: all the categories of Wikipedia articles that are not present in the dataset are ignored while computing the percentage of official categories present in the top 20 categories.



Amount of Wikipedia articles for which at least a given percentage of official Wikipedia categories was present in the top 20 categories. The "distmax=n" curves represent the case when instead of the official category we also accept one of its sub- or super-categories, assuming the level of indirection does not exceed n.

Conclusions

- *Thanks to the experimental setting choice and despite the small size of our dataset and the reduced set of categories, this method is able to predict the original categories of a document quite well.*