allahyari_2017_a_knowledge_based_topic_modeling_approach_for_automatic_topic_labeling

Year

2017

Author(s)

Mehdi Allahyari and Seyedamin Pouriyeh and Krys Kochut and Hamid Reza Arabnia

Title

A Knowledge-based Topic Modeling Approach for Automatic Topic Labeling

Venue

International Journal of Advanced Computer Science and Applications

---

Topic labeling

Focus

Type of contribution

Underlying technique

Topic labeling parameters

Label generation

In the proposed model, we define another latent (i.e. hidden) variable called, concept, between topics and words.
Thus, each document is a mixture of topics, while each topic is made up of concepts, and finally, each concept is a probability distribution over the vocabulary.

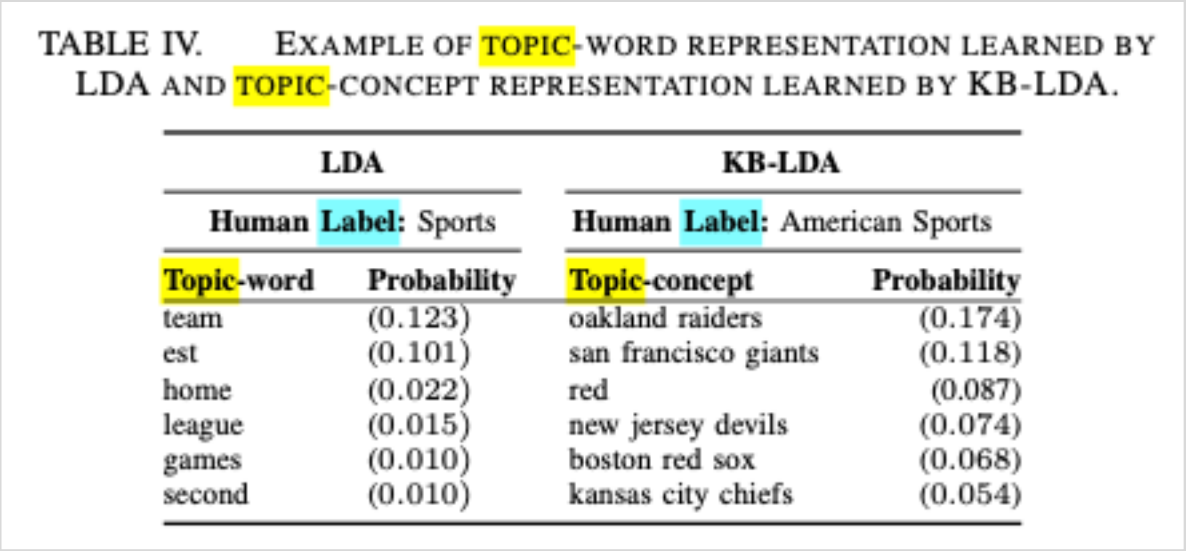


TABLE IV. EXAMPLE OF TOPIC-WORD REPRESENTATION LEARNED BY LDA AND TOPIC-CONCEPT REPRESENTATION LEARNED BY KB-LDA.

LDA		**KB-LDA**	
Human Label: Sports		**Human Label:** American Sports	
Topic-word	**Probability**	**Topic-concept**	**Probability**
team	(0.123)	oakland raiders	(0.174)
est	(0.101)	san francisco giants	(0.118)
home	(0.022)	red	(0.087)
league	(0.015)	new jersey devils	(0.074)
games	(0.010)	boston red sox	(0.068)
second	(0.010)	kansas city chiefs	(0.054)

We define a labeling approach for topics considering the semantics of the concepts that are included in the learned topics in addition to existing ontological relationships between the concepts of the ontology.

In other words, our aim is to use the semantic knowledge graph of concepts in an ontology (e.g., DBpedia) and their diverse relationships with unsupervised probabilistic topic models (i.e. LDA), in a principled manner and exploit this information to automatically generate meaningful topic labels.

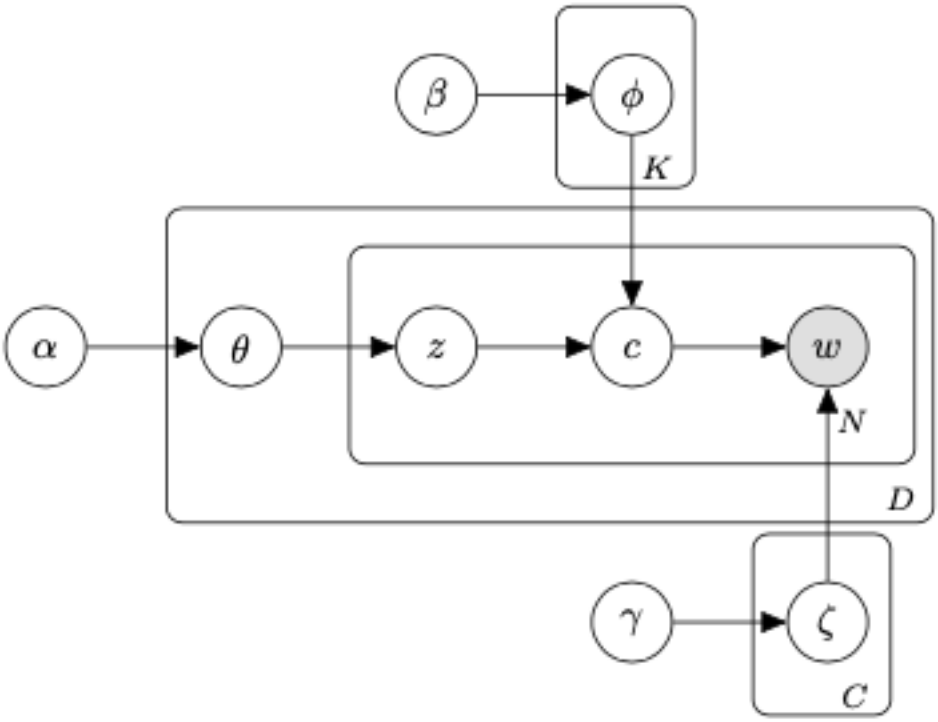
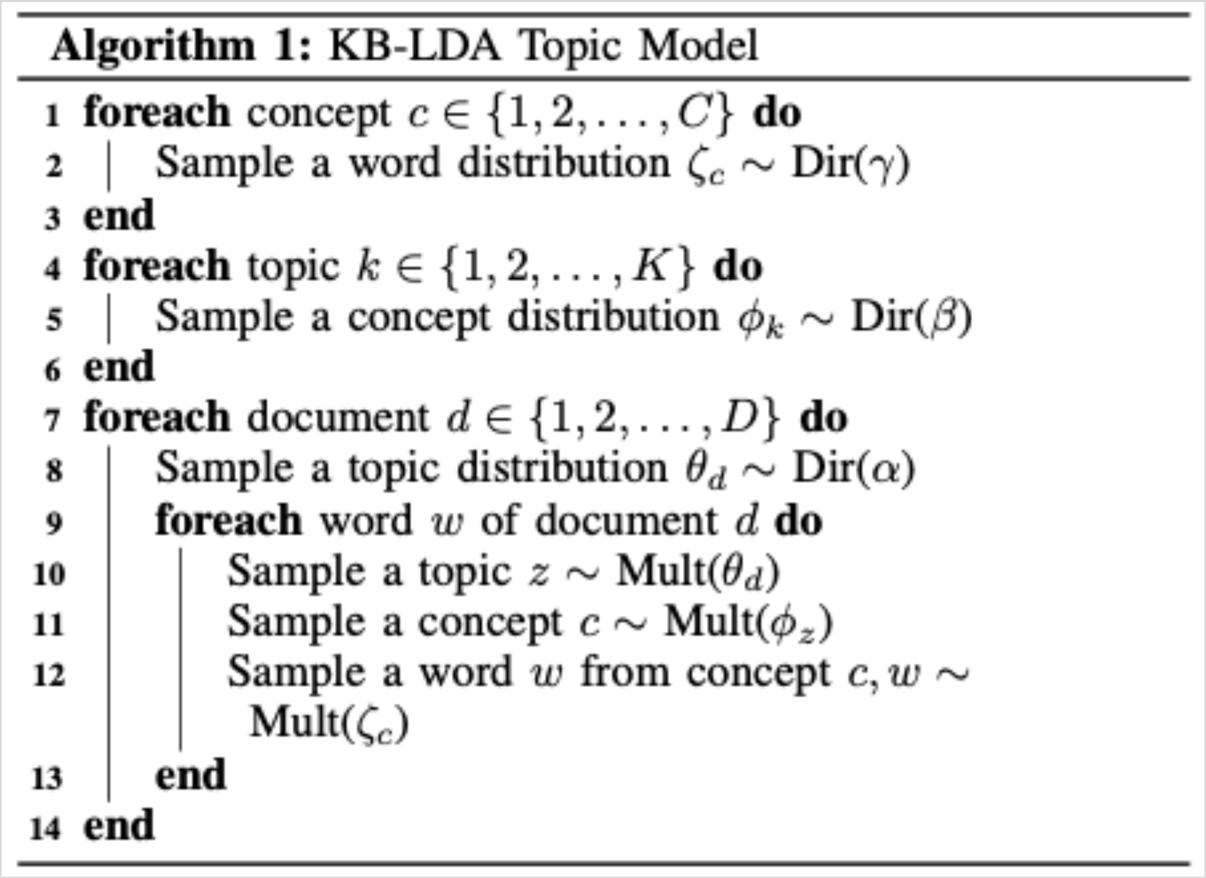


Fig. 2. Graphical representation of KB-LDA model.

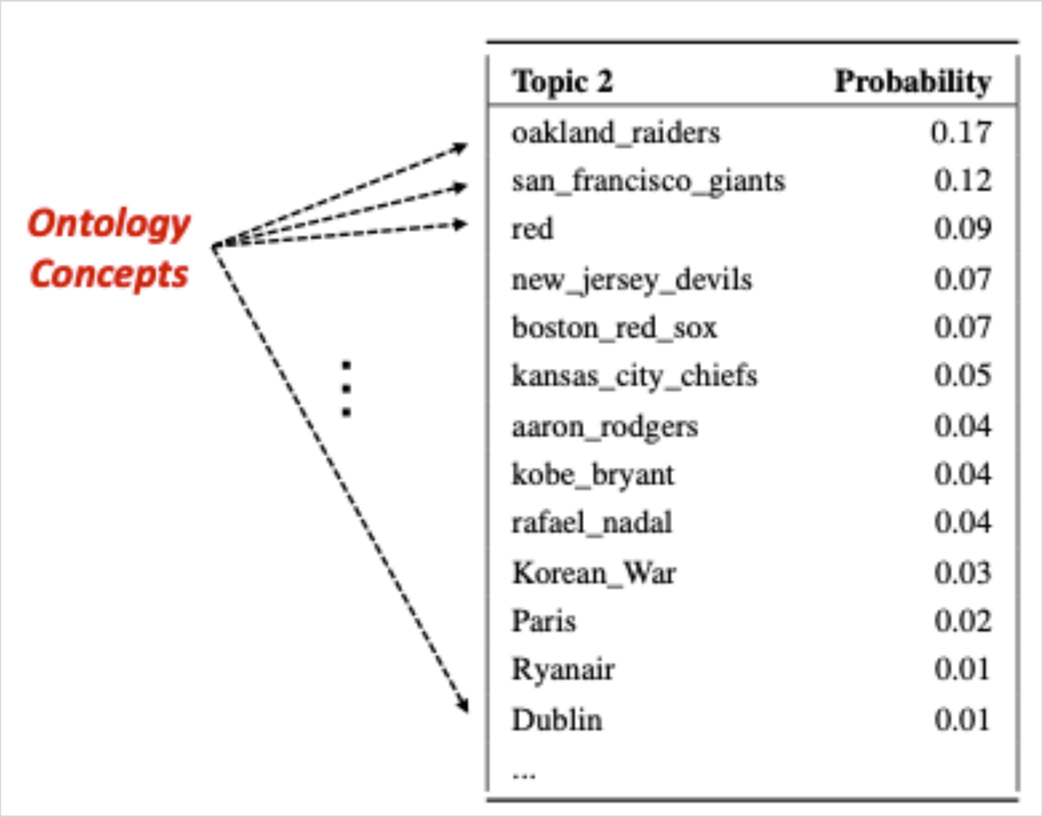


Algorithm 1: KB-LDA Topic Model

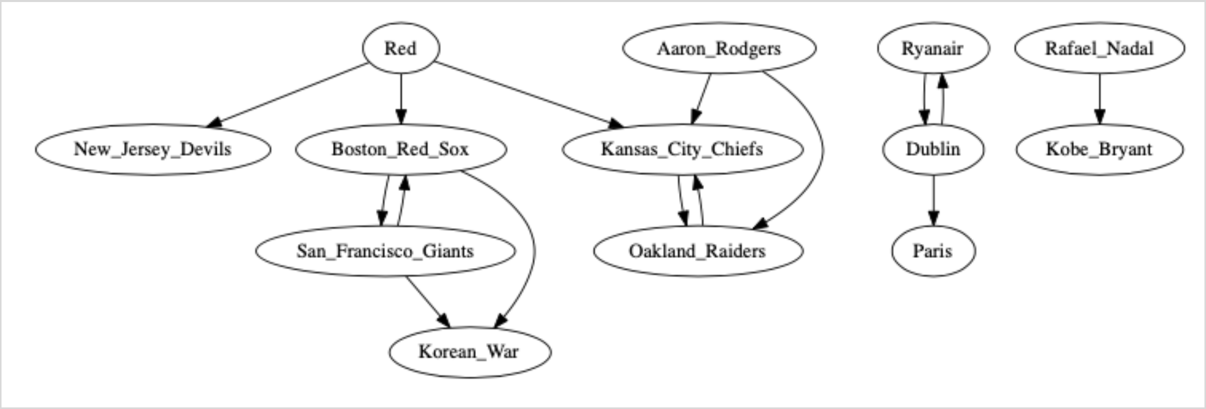
```
foreach concept $c \in \{1, 2, \ldots, C\}$ do
    Sample a word distribution $\zeta_c \sim \mathrm{Dir}(\gamma)$
end
foreach topic $k \in \{1, 2, \ldots, K\}$ do
    Sample a concept distribution $\phi_k \sim \mathrm{Dir}(\beta)$
end
foreach document $d \in \{1, 2, \ldots, D\}$ do
    Sample a topic distribution $\theta_d \sim \mathrm{Dir}(\alpha)$
    foreach word $w$ of document $d$ do
        Sample a topic $z \sim \mathrm{Mult}(\theta_d)$
        Sample a concept $c \sim \mathrm{Mult}(\phi_z)$
        Sample a word $w$ from concept $c, w \sim$
          $\mathrm{Mult}(\zeta_c)$
    end
end
```

CONCEPT-BASED TOPIC LABELING

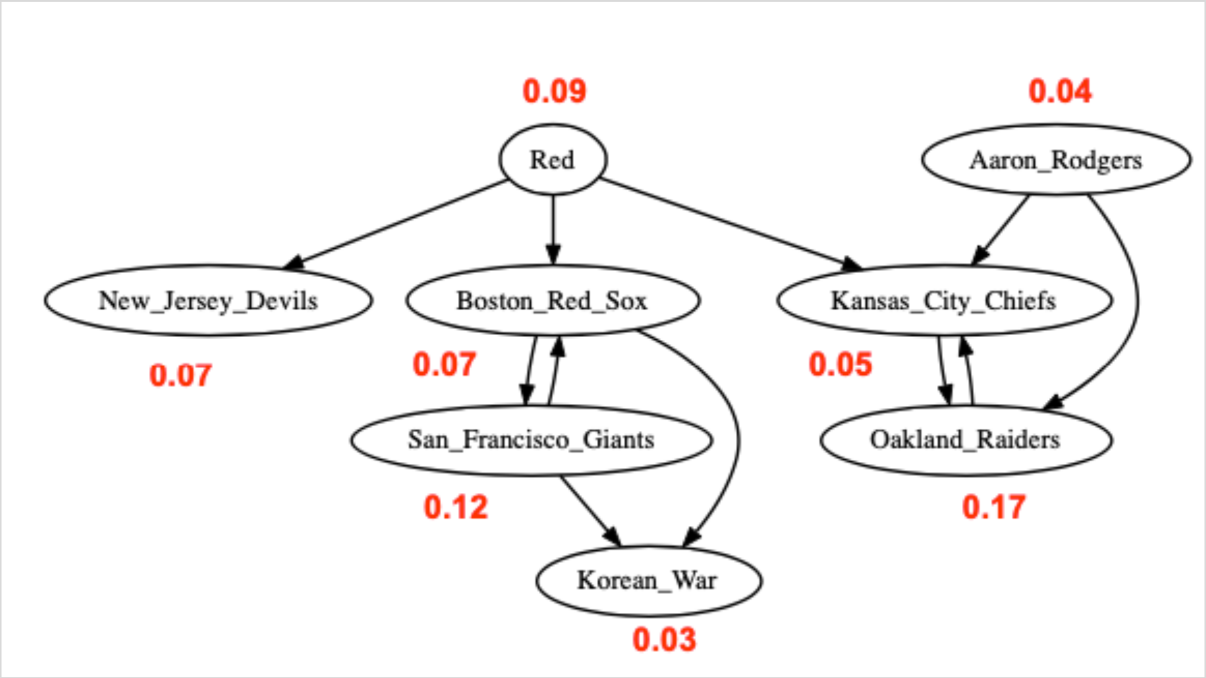
1. constructs the semantic graph from top concepts from topic-concept distribution for the given topic;



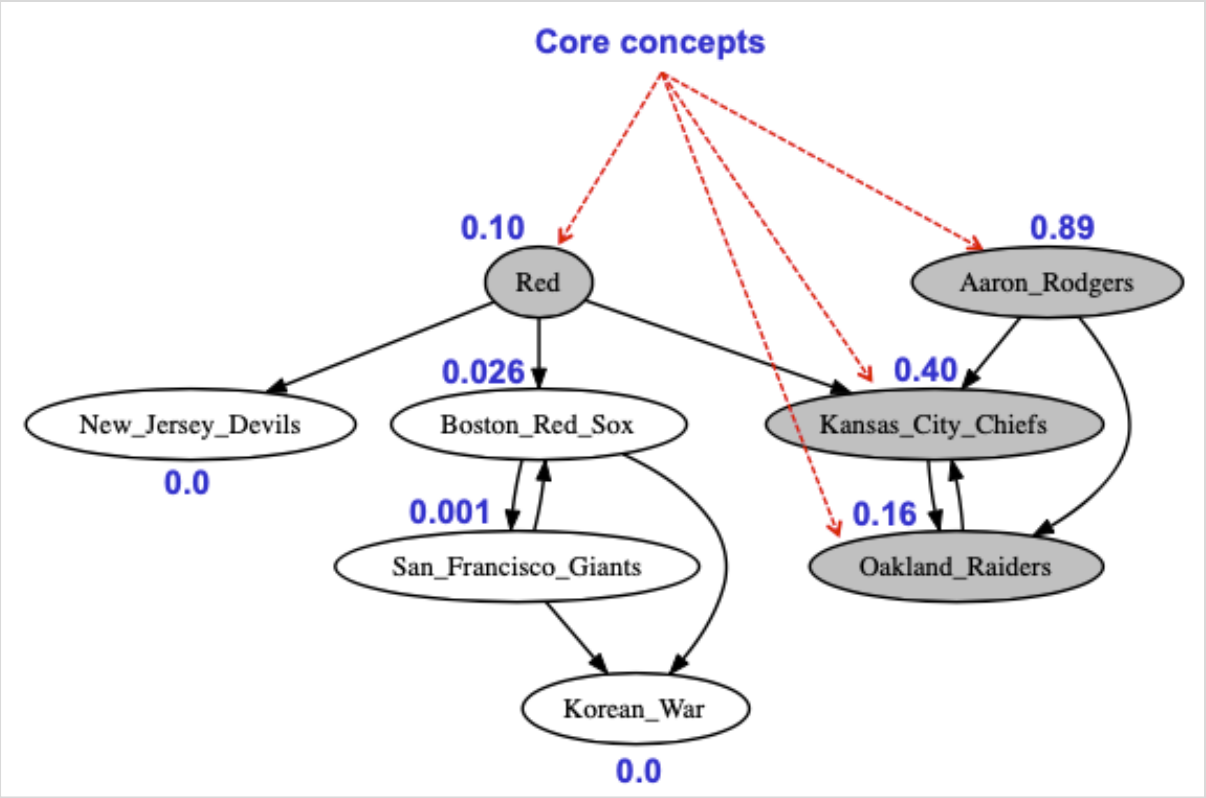
Topic 2	Probability
oakland_raiders	0.17
san_francisco_giants	0.12
red	0.09
new_jersey_devils	0.07
boston_red_sox	0.07
kansas_city_chiefs	0.05
aaron_rodgers	0.04
kobe_bryant	0.04
rafael_nadal	0.04
Korean_War	0.03
Paris	0.02
Ryanair	0.01
Dublin	0.01
...	



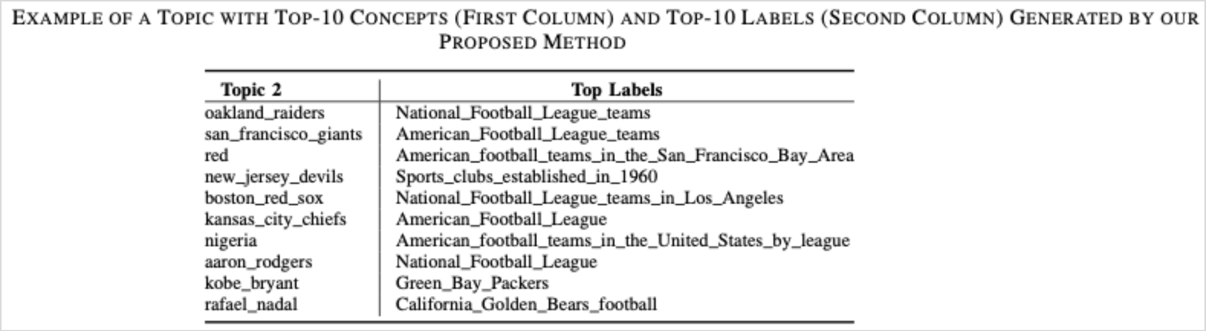
2. selects and analyzes the dominant thematic graph, a semantic graph's subgraph;



3. Extract the set of the the most authoritative and central (core) concepts in the dominant thematic graph



4. Extracts the topic label graph from the core thematic graph concepts
    1. Extract the topic label graph by traversing the ontology from each core concept and retrieving all the nodes laying at most three hops away from the core ones.
5. Computes the semantic similarity between topic and the candidate labels of the topic label graph.


EXAMPLE OF A TOPIC WITH TOP-10 CONCEPTS (FIRST COLUMN) AND TOP-10 LABELS (SECOND COLUMN) GENERATED BY OUR PROPOSED METHOD

Topic 2	Top Labels
oakland_raiders	National_Football_League_teams
san_francisco_giants	American_Football_League_teams
red	American_football_teams_in_the_San_Francisco_Bay_Area
new_jersey_devils	Sports_clubs_established_in_1960
boston_red_sox	National_Football_League_teams_in_Los_Angeles
kansas_city_chiefs	American_Football_League
nigeria	American_football_teams_in_the_United_States_by_league
aaron_rodgers	National_Football_League
kobe_bryant	Green_Bay_Packers
rafael_nadal	California_Golden_Bears_football

Motivation

Addressing the fact that:
"interpreting the label of the topics based on the distributions of words derived from the text collection is a challenging task for the users and it becomes worse when they do not have a good knowledge of the domain of the documents. Usually, it is not easy to answer questions such as "What is a topic describing?" and "What is a representative label for a topic?""

Additionally, using ontological concepts:

“as an extra latent variable (i.e. represent- ing topics over concepts instead of words) are advantageous in several ways including: (1) it describes topics in a more extensive way; (2) it also allows to define more specific topics according to ontological concepts, which can be eventually used to generate labels for topics; (3) it automatically incorporates topics learned from the corpus with knowledge bases.”

Topic modeling

Topic modeling parameters

Nr. of topics

Label

Label selection

Label quality evaluation

Assessors

Domain

Paper:
Dataset:

Problem statement

In this paper, we are taking concepts of ontology into consideration instead of words alone to improve the quality of generated labels for each topic.

We have highlighted some aspects of our approach including:

1. we have incorporated ontology concepts with statistical topic modeling in a unified framework, where each topic is a multinomial probability distribution over the concepts and each concept is represented as a distribution over words
2. a topic labeling model according to the meaning of the concepts of the ontology included in the learned topics. The best topic labels are selected with respect to the semantic similarity of the concepts and their ontological categorizations.

We demonstrate the effectiveness of considering ontological concepts as richer aspects between topics and words by comprehensive experiments on two different data sets.

Corpus

Origin:
Nr. of documents:
Details:

Document

Pre-processing

```
@article{2017_allahyari_a_knowledge_based_topic_modeling_approach_for_automatic_topic_labeling,
  author = {Mehdi Allahyari and Seyedamin Pouriyeh and Krys Kochut and Hamid Reza Arabnia},
  date-added = {2023-03-30 17:29:16 +0200},
  date-modified = {2023-03-30 17:29:16 +0200},
  doi = {10.14569/IJACSA.2017.080947},
  journal = {International Journal of Advanced Computer Science and Applications},
  number = {9},
  publisher = {The Science and Information Organization},
  title = {A Knowledge-based Topic Modeling Approach for Automatic Topic Labeling},
  url = {http://dx.doi.org/10.14569/IJACSA.2017.080947},
  volume = {8},
  year = {2017}}
```

#Thesis/Papers/BS