Recommender System: Retrieving Wikipedia links

Asmaa Demny, Fatima Habib, Cécile Macaire,
Chanoudom Prach, Ludivine Robert

UE 901 EC3

M2 NLP - University of Lorraine, IDMC

February 15, 2021

Table of contents

Introduction

Introduction

- Recommender systems can be defined as filtering system that seeks to predict the "rating" or "preference" a user would give to an item.
- The purpose is to apply recommendation systems techniques on Wikipedia dataset to output interesting suggestions.
- Generating text or information is one of the NLP tasks that can be done by text and data analysis.

Dataset

General Information

ENWIKI DUMP PROGRESS ON 20210101[1] → Wikimedia dump XML file of 166.3 MB.

Main information:

- the **title** (<title>),
- the **ID** (<id>),
- the **parent ID** (<parentid>),
- the **contributor** (<contributor>),
- the **text** (<text>) which contains, for some pages the infobox, the different sections identified by '==History==', and references (<ref>).

[1] https://dumps.wikimedia.org/enwiki/20210101/

Page Example


```
  <page>
    <title>Colegio de Santa Cruz de Tlatelolco</title>
    <ns>0</ns>
    <id>8554864</id>
    <revision>
      <id>965714004</id>
      <parentid>934400346</parentid>
      <timestamp>2020-07-03T00:06:35Z</timestamp>
      <contributor>
        <username>GreaterPonce665</username>
        <id>30826712</id>
      </contributor>
      <minor />
      <comment>{{start date and age}}</comment>
      <model>wikitext</model>
      <format>text/x-wiki</format>
      <text bytes="16412" xml:space="preserve">{{Infobox university
|name              = Colegio de Santa Cruz de Tlatelolco
|image =
File:Iglesia_de_Santiago_Tlatelolco,_M%C3%A9xico_D.F.,_M%C3%A9xico,_2013-10-16,_DD_38.JPG
|native_name       =
|motto             =
|established       = {{start date and age|1536|1|6}}
|type              = [[Catholic education|Catholic]]
|city              = [[Tlatelolco (Mexico City)|Tlatelolco]], [[Mexico City]]
|country           = [[Mexico]]
|campus            = [[urban area|Urban]]
}}
[[File:Iglesia de Santiago Tlatelolco, México D.F., México, 2013-10-16, DD 31.JPG|thumbnail|
Exterior of the church]]
[[File:Iglesia de Santiago Tlatelolco, México D.F., México, 2013-10-16, DD 46.JPG|thumb|
View of dome from below]]
The '''Colegio de Santa Cruz''' in [[Tlatelolco (Mexico City)|Tlatelolco]], [[Mexico
City]], is the first and oldest European school of [[higher learning]] in the
[[Americas]]&lt;ref&gt;{{cite book|url=https://catalog.hathitrust.org/Record/101392426|
title=The first college in America: Santa Cruz de Tlatelolco.|location=Washington DC|
year=1936|author1=Steck|author2=Francis Borgia}}&lt;/ref&gt; and the first major school of
interpreters and translators in the [[New World]].&lt;ref&gt;{{cite book|chapter-
```


Figure 1: Extract from the XML file of a Wikipedia page.

Goal

This project aims to identify the existing approaches and build a recommender system that will take a Wikipedia page as an input, and will recommend 10 new links based on information that will be defined.

Methodology

Strategy

→ Content-based filtering.

- Learns a preference model which is based on a feature-based representation of the content of recommendable items [1].
- Recommends any similar items that are based on specific notation of the domain or the content of the item.

Approaches

Approaches

1. **Doc2Vec** [2]: obtains content-based representations of document data to a vector space model.

2. **TF-IDF**: statistical measure intended to evaluate how a word is relevant to a document among a collection of documents.

User interface

→ Retrieve the request of the user.

```
Please enter a Wikipedia page name:
```

```
Please enter a Wikipedia page name: https://en.wikipedia.org/wiki/Love_to_Love
Original title : Love_to_Love
Title for searching : Love to Love
Correct Wikipedia page name, we will propose you 10 related pages!
```

Figure 2: User interface when the request is correct.

User interface (2)

```
Please enter a Wikipedia page name: https://en.wikipedia.org/wiki/Impractical_Jokers
Original title : Impractical_Jokers
Title for searching : Impractical Jokers

Incorrect Wikipedia page, please retry!

Some suggestions :)

1. Impractical joker (garfield)
https://en.wikipedia.org/wiki/Impractical_joker_(garfield)
2. Impractical joker (garfield)
https://en.wikipedia.org/wiki/Impractical_joker_(garfield)
3. The impractical joker garfield and friends
https://en.wikipedia.org/wiki/The_impractical_joker_garfield_and_friends
4. The impractical joker garfield and friends
https://en.wikipedia.org/wiki/The_impractical_joker_garfield_and_friends
5. The impractical joker garfield & friends
https://en.wikipedia.org/wiki/The_impractical_joker_garfield_&_friends

Please enter a Wikipedia page name:
```

Figure 3: User interface when the request incorrect.

After the user enters a correct link, the interface will recommend different links depending on the chosen approach.

Doc2Vec

Doc2Vec - algorithms based on word2vec which represents word in a vector space model.

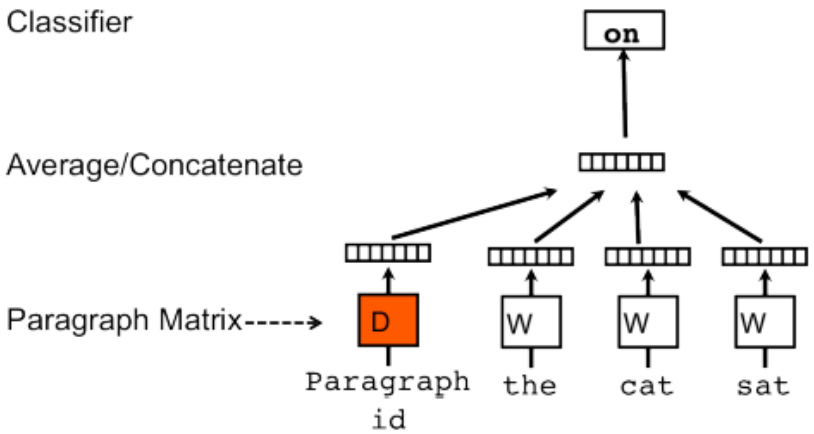


Figure 4: PV-DM model

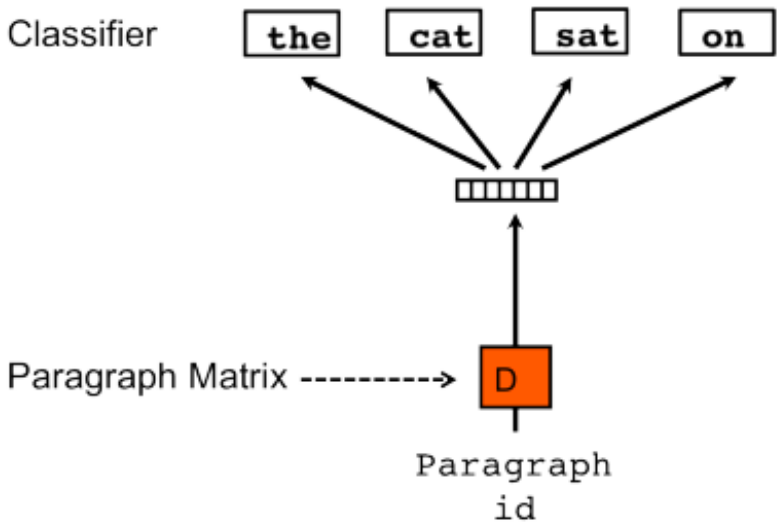


Figure 5: PV-DBOW model

Doc2Vec

Steps:

1. Load the Wikipedia dataset.
2. Run the models.
3. Generate the results from an input query (text).

Library: GENSIM.

Doc2Vec

The results of the model for a related page is the title and the score of the vectors.

```
['love', 'to', 'love']
[('Love to Love', 0.7007774114608765),
 ('The Romance of Kenny G', 0.6808857917785645),
 ('Just Like Heaven', 0.6676369905471802),
 ('Comfort and Joy', 0.6651524901390076),
 ('Sour (album)', 0.6620450019836426),
 ('Talking in Your Sleep', 0.6607742309570312),
 ('The Very Best of Kenny G', 0.6546471118927002),
 ('The Collection (Kenny G album)', 0.6523045301437378),
 ('Meet You There (album)', 0.6512295603752136),
 ('Soldier (Neil Young song)', 0.6511427760124207)]
```

Figure 6: Output of Doc2Vec model.

TF-IDF

TF-IDF - standard technique in information retrieval.

- **TF**: the frequency of a term in a document.
  $\rightarrow$ TF(t, d) = frequency of t in d / maximal frequency of a term in d.
- **IDF**: how often a term appears in all documents.
  $\rightarrow$ IDF(t) = $\log(N/n_t)$ with N, the number of all documents & n_t, the number of documents containing t.

TFIDF(t, d) = TF(t, d) * IDF(t)

TF-IDF

Several steps:

1. Extraction of the data from XML file and store them in a dataframe.
2. Preprocessing of the data.
3. Computation of the TF-IDF matrix.
4. Calculation of the cosine similarity from the TF-IDF matrix.
5. Recommendation of 10 links from the user request based on the cosine similarity scores.

Libraries: SCIKIT-LEARN, PICKLE, NLTK & REGEX.

TF-IDF: Extraction of the data

	Title	ID	Text
0	Chestnut Ridge Middle School	8554860	#REDIRECT[[Washington Township Public School D...
1	Colegio de Santa Cruz de Tlatelolco	8554864	{{Infobox university\n|name = Col...
2	Impractical joker (garfield)	8554867	#REDIRECT [[List of Garfield and Friends episo...
3	National Council of Teachers	8554873	'''National Council of Teachers''' may refer t...
4	Shuo Wang	8554878	#REDIRECT [[Wang Shuo]]
5	The impractical joker garfield and friends	8554883	#REDIRECT [[List of Garfield and Friends episo...
6	Order of battle at Beiping–Tianjin	8554884	'''Peiking Tientsin Operation''' (July–August ...
7	Gulshani	8554885	{{about|the Sufi order|the demonym of Gulshan|...
8	The impractical joker garfield & friends	8554892	#REDIRECT [[List of Garfield and Friends episo...
9	The impractical joker garfield	8554898	#REDIRECT [[List of Garfield and Friends episo...

Figure 7: Beginning of the dataframe with title, id and text information for each page.

TF-IDF: Data Preparation

Preprocessing on text:

- Remove HTML tags.
- Remove URLs.
- Remove punctuation marks.
- Remove stop words.
- Remove numbers.

TF-IDF: Matrix

Parameter	Value
analyser	word
ngram_range	(1,2)
min_df	0
max_features	1000
stop_words	english

Table 1: TfidfVectorizer parameters.

Input size: 130000.

TF-IDF: Cosine similarity

Cosine similarity – measures the cosine of the angle between two vectors projected in a multi-dimensional space.

$$similarity(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| * \|\vec{b}\|}$$

Coding line: cosine_similarities = cosine_similarity(tfidf, tfidf[:50000])

```
[[1.         0.08876145 0.         ... 0.         0.         0.        ]
 [0.08876145 1.         0.00862954 ... 0.09056698 0.         0.        ]
 [0.         0.00862954 1.         ... 0.02604671 0.05420052 0.16458987]
 ...
 [0.         0.02352993 0.         ... 0.06226215 0.         0.        ]
 [0.         0.01318819 0.01133071 ... 0.00997876 0.         0.        ]
 [0.01710205 0.03296025 0.0090965  ... 0.00222621 0.         0.        ]]
```

Figure 8: Cosine similarity matrix.

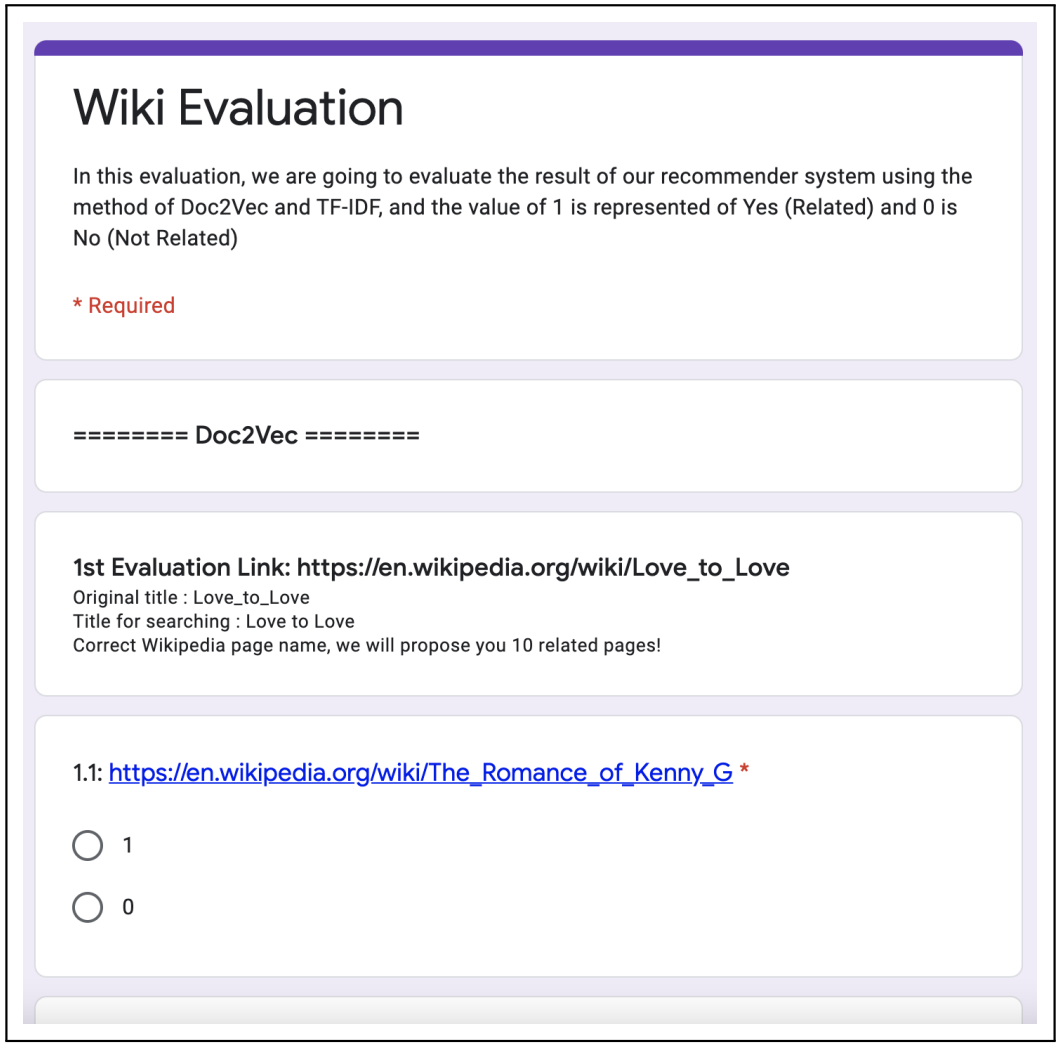
TF-IDF: Recommendations

```
Request :  https://en.wikipedia.org/wiki/Love_to_Love
Recommending 10 links similar to Love to Love page...

1. https://en.wikipedia.org/wiki/Pjetër_Dungu (score:0.083)

2. https://en.wikipedia.org/wiki/New_York_State_Route_52_Business (score:0.061)

3. https://en.wikipedia.org/wiki/Good_Night,_Little_Ones! (score:0.046)

4. https://en.wikipedia.org/wiki/The_Silence_of_the_Lambs (score:0.045)

5. https://en.wikipedia.org/wiki/Criminal_court (score:0.024)

6. https://en.wikipedia.org/wiki/1982_Topps (score:0.009)

7. https://en.wikipedia.org/wiki/Gene_D._Block (score:0.008)

8. https://en.wikipedia.org/wiki/Judicial_intern (score:0.008)

9. https://en.wikipedia.org/wiki/Colegio_de_Santa_Cruz_de_Tlatelolco (score:0.007)

10. https://en.wikipedia.org/wiki/Diarmuid_O'Neill (score:0.006)
```

Figure 9: Recommendation links for the Wikipedia page name *Love to Love* and their associated cosine similarity score.

Results & Evaluation

Evaluation



Wiki Evaluation

In this evaluation, we are going to evaluate the result of our recommender system using the method of Doc2Vec and TF-IDF, and the value of 1 is represented of Yes (Related) and 0 is No (Not Related)

* Required

======== Doc2Vec ========

1st Evaluation Link: https://en.wikipedia.org/wiki/Love_to_Love
Original title : Love_to_Love
Title for searching : Love to Love
Correct Wikipedia page name, we will propose you 10 related pages!

1.1: https://en.wikipedia.org/wiki/The_Romance_of_Kenny_G *

- 1
- 0

Figure 10: Evaluation on Google Survey Form

Evaluation (cont.)

1. Subjective evaluation.
2. Google survey form.
3. Comparison of the two models...

Results

Doc2Vec	Total Score	Average Score
1st Evaluation Link	**31**	**62%**
2nd Evaluation Link	26	52%
3rd Evaluation Link	2	4%

Figure 11: Evaluation result of Doc2Vec.

TF-IDF	Total Score	Average Score
1st Evaluation Link	5	10%
2nd Evaluation Link	**40**	**80%**
3rd Evaluation Link	14	28%

Figure 12: Evaluation result of TF-IDF.

Results (cont.)

Methods	Avg Score	Standard Deviation
Doc2Vec	39%	**31%**
TF-IDF	39%	**36%**

Figure 13: Result of Avg and Std on Doc2Vec and TF-IDF.

Discussion

Challenges

- Not a clear difference between the two approaches.
  $\rightarrow$ small amount of data,
  $\rightarrow$ evaluation limited to our group of 5 people and on 3 Wikipedia pages.

- Disadvantages from TFIDF:
  $\rightarrow$ based on the BoW model,
  $\rightarrow$ can not capture the semantic information.

- Computation and memory limitations.
  $\rightarrow$ Parallelism of jobs.

Future work

- Improve the algorithms by increasing the computational capacity.
- Use the parallelism technique in code ruining on TF_IDF.
- Test different parameters to get more tune-fining results.
- Implement approaches based on deep learning techniques.

References

[1] P. Lops, D. Jannach, C. Musto, T. Bogers, and M. Koolen, "Trends in content-based recommendation," *User Modeling and User-Adapted Interaction*, vol. 29, no. 2, pp. 239–249, 2019.

[2] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, PMLR, 2014, pp. 1188–1196.

Thank you!

Do you have any questions?