Big Data Project-2
Sentence Similarity on CORD Dataset

By

Yash Kasundra

Masters of Data Science

Outline

- Introduction
- Methodology
- Implementation
- Results and Findings

Introduction

- Finding top 10 most similar articles and sentences from those articles using input questions from users:
  - To begin with, we have to create a Kaggle kernel to read all the Jason files and create an csv file for getting a smaller subset of the dataset.
  - Dataset: Obtained from Kaggle website
  - Goal: To build or find a good model which can find similar sentences from articles based on the input, using cosine score as similarity index.

- Property:
  - Sentence similarity NLP problem
  - There are 2 columns “abstract and body_text” which we will be using mainly.
  - We are to make some features from this column in order to get good accuracy
  - Check number of articles in file generated from 1st notebook.

Methodology

- Since we need to find sentence similarity we will be using mainly 3 models mentioned below:
  - Word 2 Vector (word2vec)
  - Document 2 Vector (doc2vec)
  - Sentence transformation (sbert)

Implementation

1. Exploratory Data Analysis (EDA)

2. Pre-processing the data to clean unwanted text

3. Modeling :-
    1. Loading/Training 3 models
    2. Taking user's input
    3. Passing into word2vec first.
    4. Then giving user a choice to select among 2 doc2vec model
        1. Trained on abstract
        2. Trained on body_text
    5. Implemented sentence transformation to get 10 most similar articles

1) Exploratory Data Analysis (EDA)

- Check if any Null or duplicate values.

```
df.isnull().sum()
```

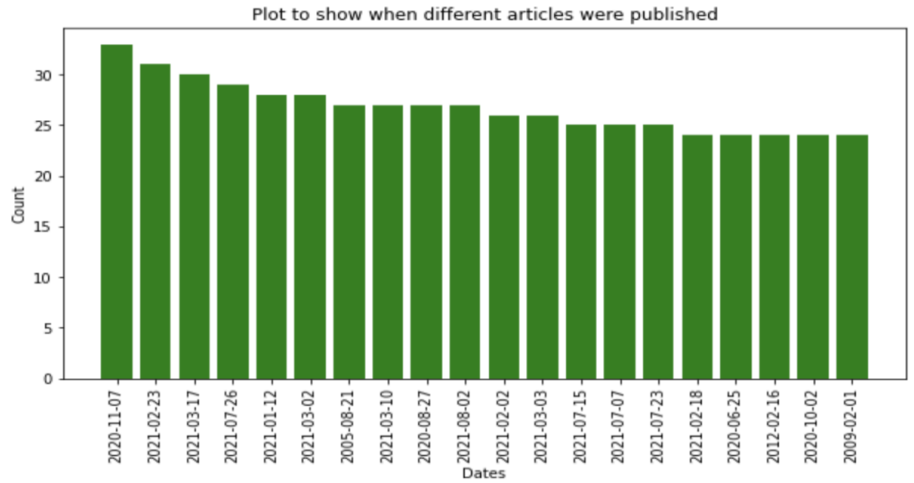
```
cord_uid          0
title             0
abstract          0
publish_time      0
authors           0
journal           0
pdf_json_files    0
pmc_json_files    0
url               0
body_text         0
dtype: int64
```

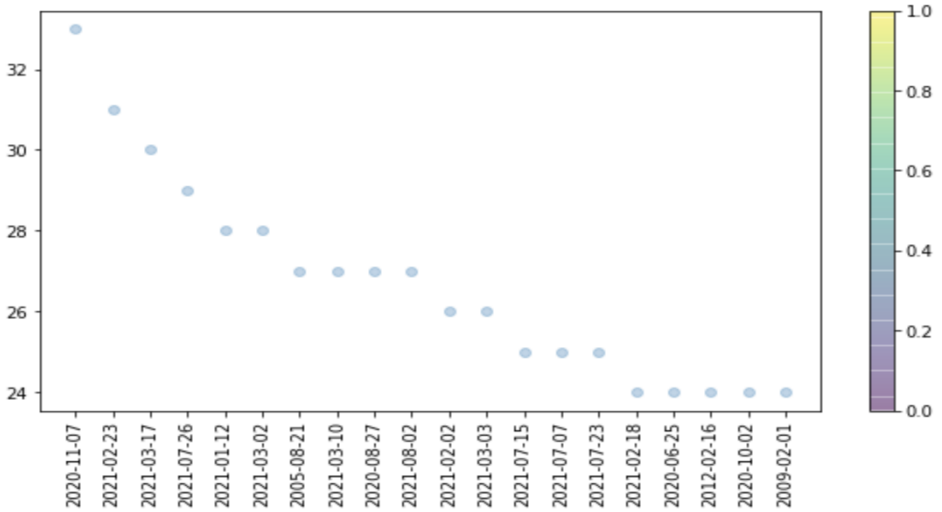
hecking if there are any duplicates present

```
df.duplicated(subset={"title","abstract","body_text"}).value_counts()
```

```
False    7985
dtype: int64
```

- Plots to see number of publications according to time.

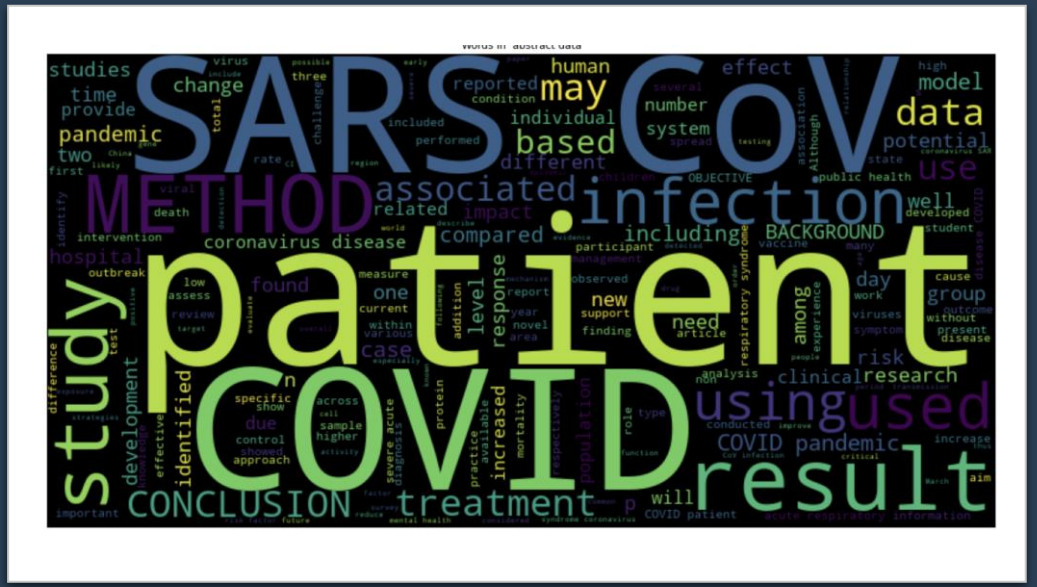


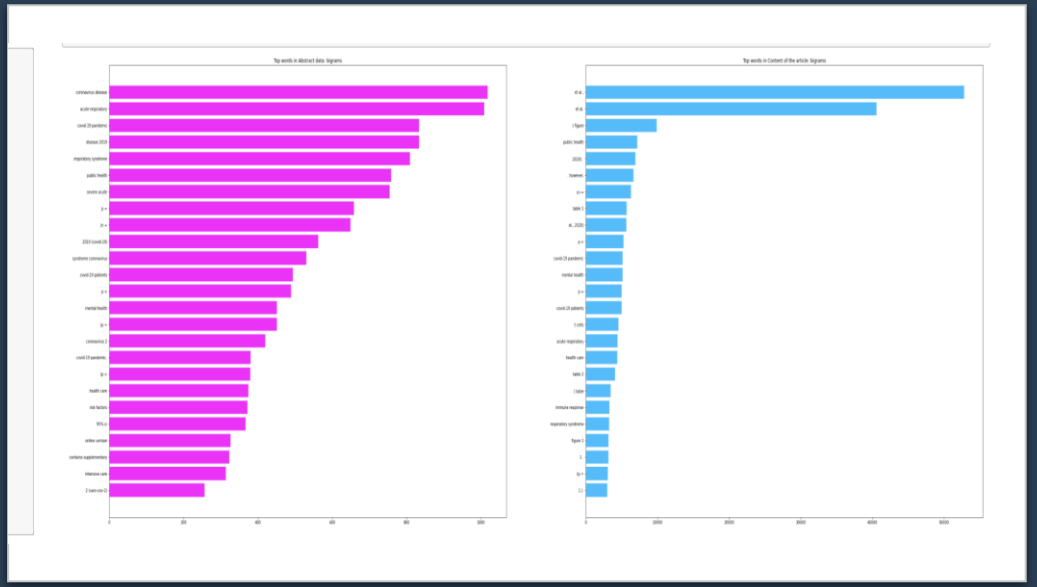


- Wordclouds to see most used words. This first image is an example of word cloud created form abstract.

- N-gram frequency plots to see word combos used most. This is a frequency plot for set of 2 co-occurring used words in our data.
  - There are 2 more plots with 1-word and 3-words frequency plot.

- Created a function to check all the special characters and symbols present in abstract/ body_text columns. This the output for using the function on body_text





2) Data Pre-processing

- I created few function for :
  - Replacing math equations and url's with text “MATH EQUATION” and “URL”.
  - Cleaning contractions. ("We'd": "We had“ , "O'Clock": "Of the clock")
  - Spelling Correction. ('organisation': 'organization’, 'cryptocoin':'bitcoin’)
  - Removing punctuations. ('@', '£', '·', '_', '{‘)
  - Removing Stopwords. (‘is’, ’a’, ’an’)
  - Using WordNet Lemmatizer (“Causing” : “cause” , “hunted” : “hunt”)
  - And few functions to convert whole articles into list of sentences.

- Then again visualize after preprocessing, to gain insights.

```
In [66]: print(df['body_text'][0])
```

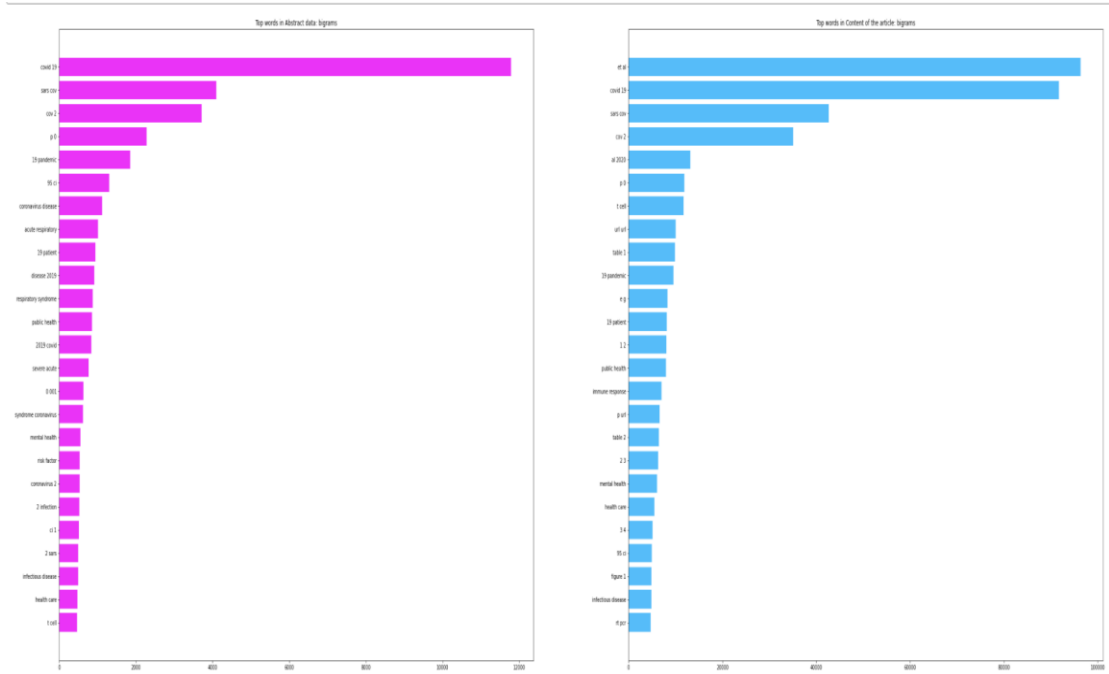
```
Introduction: Setting the Scene

In the past decade, Ireland has experienced an intense debate about the suitability of the structure of the primary educa
tion sector to meet the needs of an increasingly diverse, multi-faith society. The primary system is currently denominati
onal and diverse, with schools under Catholic patronage occupying the largest sector (88.9%). It should be acknowledged t
hat the system evolved from the historical development of the country, and reflects the unique Church/State relations tha
t previously existed. It is also important to note that Ireland is still largely a Christian country, with the majority o
f the population identifying as Catholic in the most recent census.The challenges facing Catholic education globally are
well documented, as evidenced in the increased interest in scholarship and research in the field. Ireland is not immune t
o these challenges, with an ongoing questioning of the role and value of faith schools in a secular society. Notably, the
re is an intensified focus on the subject area of Religious Education in these schools due to high number of Catholic sch
ools at primary level. It could be argued that it is this unique structural and patronage context that makes the critique
about Catholic schooling in Ireland different to the experiences of other countries.The ongoing debate is fuelled by a nu
mber of recent landmark educational developments, such as the Report from the Forum on Patronage and Pluralism in the Pri
mary Sector (2012) and the Draft Primary Curriculum Framework (2020), which signal significant implications for Religious
Education in Catholic primary schools. This chapter will explore these developments in relation to Religious Education an
d the State, the educator, and the Church.

Religious Education and the State
```

```
In [67]: print(df['preprocessed_body_text'][0])
```

```
introduction setting scene in past decade ireland experienced intense debate suitability structure primary education sect
or meet need increasingly diverse multi faith society the primary system currently denominational diverse school catholic
patronage occupying largest sector 88 9 it acknowledged system evolved historical development country reflects unique chu
rch state relation previously existed it important note ireland still largely christian country majority population ident
ifying catholic recent census the challenge facing catholic education globally well documented evidenced increased intere
st scholarship research field ireland immune challenge ongoing questioning role value faith school secular society notabl
y intensified focus subject area religious education school due high number catholic school primary level it argued uniqu
e structural patronage context make critique catholic schooling ireland different experience country the ongoing debate f
uelled number recent landmark educational development report forum patronage pluralism primary sector 2012 draft primary
curriculum framework 2020 signal significant implication religious education catholic primary school this chapter will ex
plore development relation religious education state educator church religious education state one significant milestone
catholic schooling ireland report forum patronage pluralism primary sector 2012 this landmark report initially looked pro
cess divesting greater diversity choice regard patronage primary school although initial term reference made number recom
mendation potentially impact ethos characteristic spirit religious education school the process forum subsequent recommen
dation broadly welcomed education stakeholder an area proved contentious proposal introduce subject education religion be
lief erb ethic curriculum school addition subject area religious education perhaps one reason proposed new subject erb et
hic met contention uncertainty surrounded purpose namely oversight subject related patron s religious education programme
a state body national council curriculum assessment ncca responsibility designing erb ethic apparent purpose provide neut
ral subject student opting patron s religious education programme there appeared lack clarity question raised implication
```



3) Modeling

1. First, we used pretrained word2vec model.

```
# Refer this documentation for word2vec https://radimrehurek.com/gensim/models/word2vec.html
import gensim.downloader as api

wv = api.load('word2vec-google-news-300')
```

```
[==================================================] 100.0% 1662.8/1662.8MB downloaded
```

2. Took input from user

```
# Taking user's input to get the question then change it to list to pass into our "prepare_input_df" function
num = input("Please enter number of questions you want to ask.")

input_list1= []
for i in range(int(num)):
    sentence = input();
    input_list1.append(sentence)

input_list1
```

```
Please enter number of questions you want to ask.1
```

3. Passed into word2vec model to get most similar sentences from articles.

body	abstract	preprocessed_abstract	preprocessed_body_text	sentences	token_sentences	extracted_relevant_sentences_word2vec_dianostics_results
g the past...	Does Religious Education have a future? It cou...	doe religious education future it argued conte...	introduction setting scene in past decade irel...	[Introduction: Setting the Scene\n\nIn the pas...	[[introduction, setting, scene, past, decade, ...	[If education were really understood as the pr...
d den seine ...	Die genauere Analyse in diesem Kapitel zeigt, ...	die genauere analyse diesem kapitel zeigt das ...	einem land den notwendigen raum für seine wirt...	[\n\neinem Land den notwendigen Raum für seine...	[[einem, land, den, notwendigen, raum, seine, ...	[]
neous ting...	BACKGROUND: Most post-licensure vaccine pharma...	background most post licensure vaccine pharmac...	background spontaneous passive reporting syste...	[Background\n\nSpontaneous or passive reportin...	[[background, spontaneous, passive, reporting,...	[Using active surveillance systems, new vaccin...
imary peni...	Immune thrombocytopenia (ITP) is a disease of ...	immune thrombocytopenia itp disease heterogeno...	introduction primary immune thrombocytopenia i...	[Introduction\n\nPrimary immune thrombocytopen...	[[introduction, primary, immune, thrombocytope...	[This may be more applicable in at-risk adults...
Background: The						

4. Again giving user a choice to select from 2 models as shown:

```
# Giving user an chance to choose which model they want to use for predictions (one that is trained on abstract or body_text)
model_choose = input("Please enter 1) If you want to use model trained on abstarct data or 2) If you want to use model traine
```

```
Please enter 1) If you want to use model trained on abstarct data or 2) If you want to use model trained on body_text.
```

Results and Findings

5. Using the model selected by user, we get the output:-
   1. In this case user choose to use model trained on articles

	Tags Generated by Doc2Vec for articles	abstract	extracted_relevant_sentences_word2vec_dianostics_results	original_body	cosine_similarity
0	963	BACKGROUND: Vaccination is the most important ...	[Due to zoonotic nature of leptospirosis , it...	\n\na1111111111 a1111111111 a1111111111 a11111...	0.422505
1	1419	This paper reviews the application of the algo...	[Classical PSO\n\nA swarm is usually thought a...	Introduction\n\nThe link between geophysical d...	0.373506
2	6854	Physical activity and a healthy diet are key f...	[Given the possibility of new virus outbreaks ...	Introduction\n\nThe health of today's working ...	0.368589
3	2637	A mutation of just one gene will cause abnorma...	[The With the use of two MAbs a simple, reliab...	Introduction\n\ntherapy, including cell and nu...	0.367693

   2. In this case user choose to use model trained on body_text

	Tags Generated by Doc2Vec for articles	abstract	extracted_relevant_sentences_word2vec_dianostics_results	original_body	cosine_similarity
0	824	Community-acquired pneumonia (CAP) is a common...	[CFPNGS may also present a challenge for infec...	Introduction\n\nPediatric community-acquired p...	0.651747
1	7141	COVID-19 caused rapid mass infection worldwide...	[The cut-off value was 6, suggesting that an i...	Introduction\n\nIn December 2019, many cases o...	0.606977
2	6521	Author reviews digital transformation of schol...	[And it would be spread instantly if it is cat...	\n\nSince the emergence of the world wide web,...	0.562907
3	1074	Coworking spaces have received a lot of attent...	[The conversion of one source of capital into ...	Introduction\n\nSince first established in San...	0.533655
4	5524	Immunodeficient mice engrafted with human peri...	[27 previously reported that NOD/SCID β 2 micr...	INTRODUCTION\n\nHuman in vivo immune responses...	0.510974

6. I also tried to implement sentence transformation (SBERT) :-

- This is just a part of the solution.

	cosine_sim	index
0	0.730520	7789
1	0.722512	259
2	0.716628	1925
3	0.712964	7538
4	0.704850	5488
5	0.704553	7881
6	0.701113	5764
7	0.701107	2215
8	0.699934	705
9	0.695874	6605

```
-------------------------------------------0-------------------------------------------
Cosine Similarity: 0.7305201292037964

Article index: 7789

Title of the article: SARS-CoV-2 and the safety margins of cell-based biological medicinal products

Abstract: With the pandemic emergence of SARS-CoV-2, the exposure of cell substrates used for manufacturing of medicines has
become a possibility. Cell lines used in biomanufacturing were thus evaluated for their SARS-CoV-2 susceptibility, and the d
etection of SARS-CoV-2 in culture supernatants by routine adventitious virus testing of fermenter harvest tested.

Body Text: Introduction

A general concern in cell-based manufacturing of recombinant proteins, including vaccines, 44 is the potential contamination
of the cell culture with viruses, which has had severe 45 consequences for patients and manufacturers [1, 2] . With the pand
emic emergence of SARS- 46 CoV-2, the exposure of biomanufacturing cell lines to a new virus has become a possibility, and t
o safeguard biomedicines it is important to understand whether the virus is even capable



-------------------------------------------1-------------------------------------------
Cosine Similarity: 0.7225115299224854

Article index: 259

Title of the article: What can we predict about viral evolution and emergence?

Abstract: Predicting the emergence of infectious diseases has been touted as one of the most important goals of biomedical s
cience, with an array of funding schemes and research projects. However, evolutionary biology generally has a dim view of pr
ediction, and there is a danger that erroneous predictions will mean a misuse of resources and undermine public confidence.
Herein, I outline what can be realistically predicted about viral evolution and emergence, argue that any success in predict
ing what may emerge is likely to be limited, but that forecasting how viruses might evolve and spread following emergence is
more tractable. I also emphasize that a properly grounded research program in disease prediction must involve a synthesis of
ecological and genetic perspectives.

Body Text: What can we predict about viral evolution and emergence?

Edward C Holmes 1, 2 Predicting the emergence of infectious diseases has been touted as one of the most important goals of b
iomedical science, with an array of funding schemes and research projects. However, evolutionary biology generally has a dim
view of prediction, and there is a danger that erroneous predictions will mean a misuse of resources and undermine public co
nfidence. Herein, I outline what can be realistically predicted about viral evolution and emergence, argue that any success
in predicting what may emerge is likely to be limited, but that forecasting how viruses might evolve and spread following em
ergence is more tractable. I also emphasize that a properly grounded research program in disease prediction must involve a s
ynthesis of ecological and genetic perspectives.

Introduction

The SARS epidemic of 2003, the global spread of swineorigin H1N1 influenza in 2009, the recent appearance of new hemorrhagic
```

- As we saw in previous slides, cosine scores of body text is quite high compared to abstract one. And there is a very simple reason behind it i.e. high amount of data in body_text compared to abstract

- Another thing I observed was that training doc2vec models were quicker than training word2vec and thus I choose to use pretrained model for word2vec and train doc2vec on the corpus.

- SBERT model is also quite fast in finding the similar article, plus its cosine scores are much higher than doc2vec models.

What next ?

- To increase the accuracy of the models, we can try using more data since in this project we used only around 7900 articles and there are still thousands more available in the data set.
- We can also try to use a model that can convert non English articles into English format, since right now our model is not able to use those articles as we saw.
- Another thing we could implement is using doc2vec to get the similar sentences as well, that way we can eliminate the use of word2vec which is take high computational power as well as memory.
- Explore more models or if there are resources like computing powers and memory available we can train a CNN which have found its way into NLP in recent years.