UNIVERSIDAD
POLITÉCNICA
DE MADRID

IREI: Profile-based data retrieval project

News recommendation based on profile bio in social networks

Ostap Kharysh: `ostap.kharysh@alumnos.upm.es`

Department of Computer Science
Master EIT Digital in Data Science

April 2, 2021

Contents

1 Introduction

The last decade has opened the age of big data, in which computational power makes it possible to extract important knowledge from data for differen purposes. On of such purposes is to analyse the human profiles and offer them products and services there are most likely to engage with. This article provides on of the solution to such tasks. Here we construct the news recommender system that advises on which information articles to offer to the user based on his interests. To provide maximally objective results we extract information from the real profiles on social networks and use the dataset of the real world news.

We use TFIDF (Term Frequency-Inverse Document Frequency) and Word2Vec approaches to explore whether the news recommended are different for the selected profile generation approach and particular world news dataset and whether their is some specific features that differentiate them.

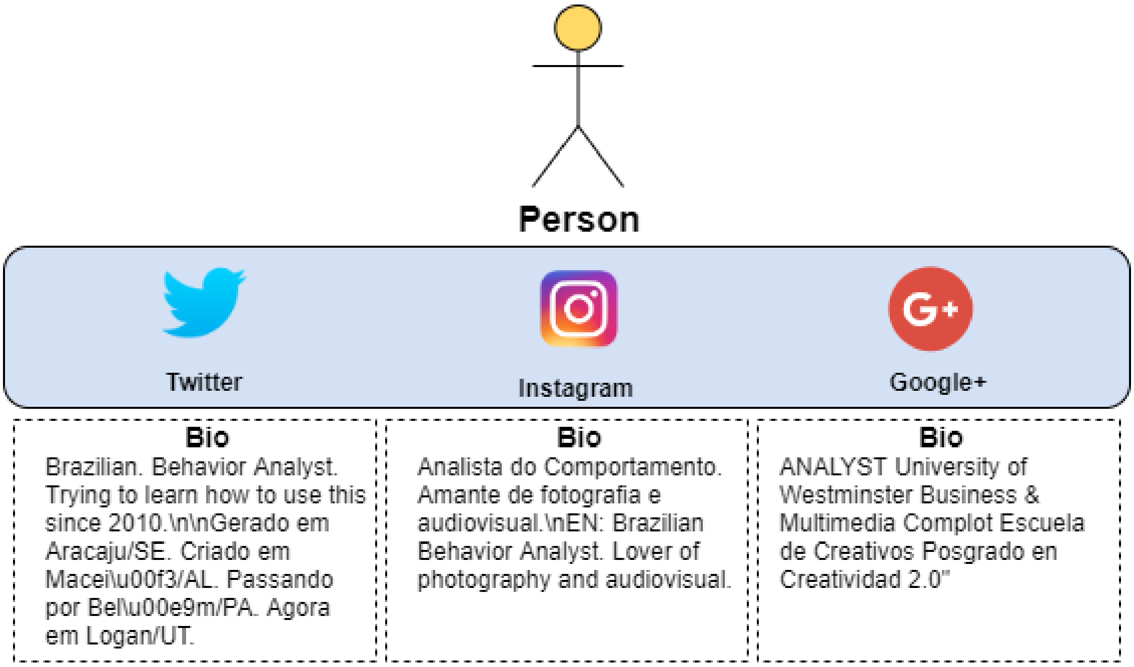
2 Data description

For this project we will use the dataset of profiles obtained from Vishal Sharma GitHub which he used for his research in ”Linking User Profiles Across Multiple Social Media Platforms” and dataset of worldnews obtained from Kaggle.

In subsections below, we will discuss applicability of those datasets for the task of news recommendation and highlight the risk associated with them.

2.1 Profile data

As we are looking to make the project as much realistic as possible we do not manually create the the profile and its topic interests but extract the the profile description of real people that they shared on their social networks accounts. The dataset of Vishal Sharma contains the profile descripiton of one person taken from Twitter, Instagram or GooglePlus. The news recommender system we describe in this report has ability to select the preferred social network from which to extract the profile information. The example of such profiles and informations in their bio could be fond in 2.1 .



Brazilian. Behavior Analyst. Trying to learn how to use this since 2010.\n\nGerado em Aracaju/SE. Criado em Macei\u00f3/AL. Passando por Bel\u00e9m/PA. Agora em Logan/UT.

Analista do Comportamento. Amante de fotografia e audiovisual.\nEN: Brazilian Behavior Analyst. Lover of photography and audiovisual.

ANALYST University of Westminster Business & Multimedia Complot Escuela de Creativos Posgrado en Creatividad 2.0"

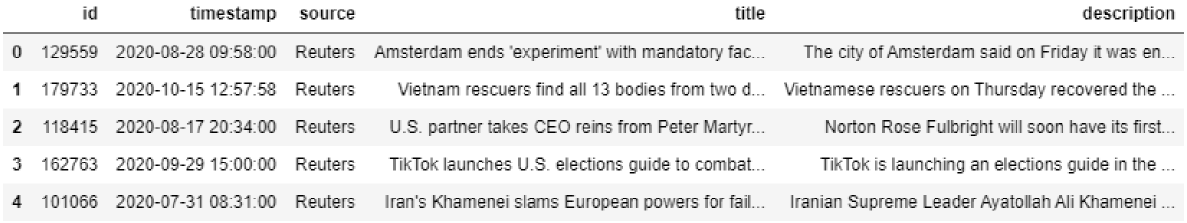
Figure 1: Example of the social networks profile information

As one could infer, there are some inconveniences this dataset could cause. For example, the could be problems with spacing, tabulations, incorrectly decoded symbols. Also some profiles are written in different languages and we will need to translate them in English as most of the profiles and news are in English.

2.2 News data

After we described how we identify profiles now we explore the news we obtained from Kaggle. The dataset contains 163,335 articles. The news are taken from such information agencies as: Foxbusiness.com, Cnet.com , The Verge, Nytimes.com ,Rawstory.com, Investors.com, Wreg.com, Reuters, Koin.com, Inc.com, CNBC, Nj.com, Wmtw.com, Nbcdfw.com, Bloomberg, Wowt.com, BBC.com. Tha data is gathered from March 21st, 2020 till January 24th, 2021.

Each news article contains: *timestamp, source, title, description.* As we are not interested in time when the news were published for our task and the agency which published it we discharge *timestamp* and *source.* For news recommendation we will use either *title* or *description* and compare which of them has higher similarity scores for a particular profile.



	id	timestamp	source	title	description
0	129559	2020-08-28 09:58:00	Reuters	Amsterdam ends 'experiment' with mandatory fac...	The city of Amsterdam said on Friday it was en...
1	179733	2020-10-15 12:57:58	Reuters	Vietnam rescuers find all 13 bodies from two d...	Vietnamese rescuers on Thursday recovered the ...
2	118415	2020-08-17 20:34:00	Reuters	U.S. partner takes CEO reins from Peter Martyr...	Norton Rose Fulbright will soon have its first...
3	162763	2020-09-29 15:00:00	Reuters	TikTok launches U.S. elections guide to combat...	TikTok is launching an elections guide in the ...
4	101066	2020-07-31 08:31:00	Reuters	Iran's Khamenei slams European powers for fail...	Iranian Supreme Leader Ayatollah Ali Khamenei ...

Figure 2: Example of news dataset

It could be possible that the news are focused mostly on general world topics as those agencies work internationally. There is a possibility that some persons could not be presented by any news closely related to their interest inferred from the profile bio in the selected social network.

In the next chapter we explain the preprocessing of the textual information we conduct to prepare the both profile information and news for the news recommendation.

3 Data preprocessing

In this phase are prerprocessing the dataset of both profiles and news before implementing conducting the similarities search for recommender system.

3.1 Profiles selection

Fist of all we select 5 random profiles of persons we would like to recommend the news. As one could infer from 2.1 the text encoding and language mix is a burden for our task so we describe how we deal with it. After we decide from which social network we want to take the persons' 'Bio' ("Twitter", "Instagram" or "GooglePlus") we randomly choose 5 people. With a help of python library google-trans-new we automatically detect whether the next is written in English, and if it's not we translate it to English. Important thing to notice here is that if the text contains words from 2 languages, for example English and Spanish, it is highly likely that it will be still identified as English. We believe, if we translate each word of the sentence we could basically end up in loosing the initial meaning the description could have in combination of words, for instance phrases or terms that are described in 2 or more words.

3.2 Text normalization

This part is conducted both for profile and news datasets. For the text normalization we will use NLTK and re.

First of all, with the help of regular expression we remove all non alphabetic values from the text, than we lower it and remove extra spacing. With the help of NLTK

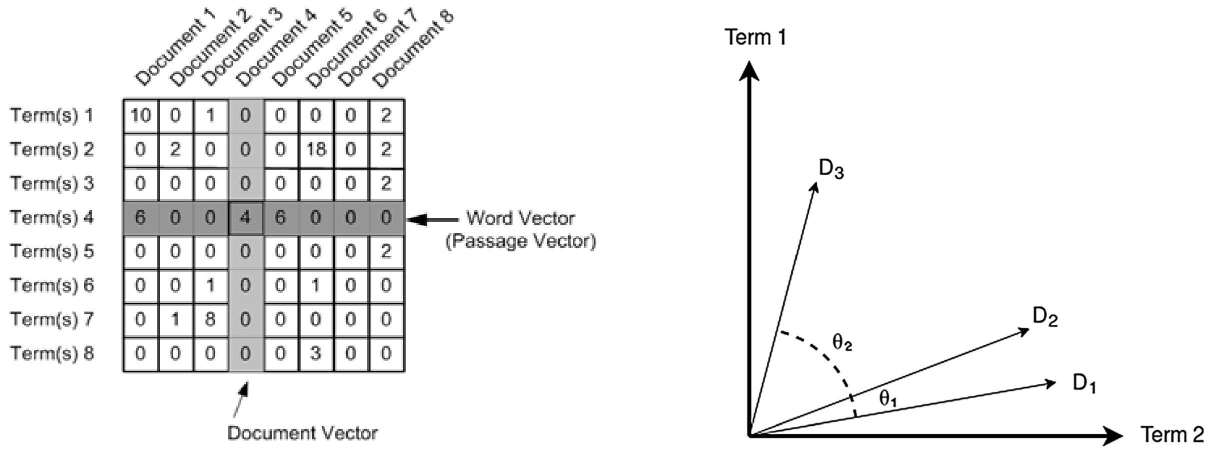
WordPunctTokenizer we create tokens of words. Each of the words is than checked whether it does not belong to the predefined stopwords in the NLTK. The ones that are considered stopwords are removed. The next phase is ability to stem or lemmatize each token and our algorithm makes this step optional for the user. By default we offer word lemmatization as we believe it's the most rational approach for our problem. With the help of lemmatization we remove inconsistencies in expraSSing the same words with different endings and exploit a major benefit of lemmatization over stemmming: considering the context to convert into a meaningful base form. For instance for a word "caring" lemmatized is "care" which contains the same meaning, but a stemmed form is "car" which loses the initial context meaning.

3.3 Text transformation

Due to memory boundaries we select randomly only 10000 news articles out of all available 163335. After we normalized them 3.2 we preprocess them for 2 approaches of recommendation: TFIDF and Word2Vec. Depending on whether we want to focus on finding similarities between the news title and profile, or news description and profile we select what we want to transform. For our purpose we selected a news description as it holds much more words and so that we won't discharge the news article which is interested to the reader that cannot be inferred from the title.

For this part we will largely use gensim and NLTK

TFIDF For TFIDF, first of all we create a dictionary of normalized texts and transform them in bag-of-words format. With the gensim implementation of TFIDF we generate the matrix of documents and terms that describes the importance of word for a given corpus (news dataset). The words from profiles' Bio are also dictionarized. Having both news articles and profile descriptions prepared we create we construct a similarity matrix of the news articles and try to find the most similar to the profiles' Bio description applying cosine similarity. The lover is the angle between the vector that represents the Bio description and news article the higher is similarity between them, hence we need to recommend such news to the person.



	Document 1	Document 2	Document 3	Document 4	Document 5	Document 6	Document 7	Document 8
Term(s) 1	10	0	1	0	0	0	0	2
Term(s) 2	0	2	0	0	0	18	0	2
Term(s) 3	0	0	0	0	0	0	0	2
Term(s) 4	6	0	0	4	6	0	0	0
Term(s) 5	0	0	0	0	0	0	0	2
Term(s) 6	0	0	1	0	0	1	0	0
Term(s) 7	0	1	8	0	0	0	0	0
Term(s) 8	0	0	0	0	0	3	0	0

Figure 3: Cosine similarity based on TFIDF

Word2Vec Now we look at the second approach which is similarities based on Word2Vec. Word2Vec is an algorithm for distributed representations of words as vectors. Also it is important to mention that Wor2Vec is a pretrained neural network model on huge dataset of diffenrent sources of textual information. So this approach allows us to take advantage of the words closeness based on the knowledge obtained through learning. Now we have a model that will distribute each word of the corpus to the same space where distances in describe closeness of terms. Finally we create a matrix of words applying the Word2Vec knowledge and than apply the person's Bio and which word vector spaces are the closest with cosine similarity which follows the same logic as in 3.3 .

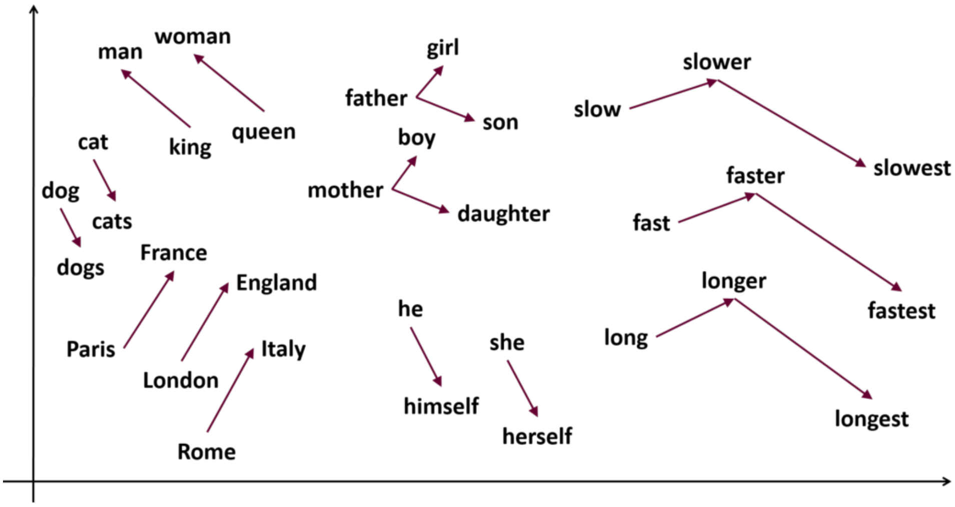


Figure 4: Example of words representation in Word2Vec

4 Results

To sum up, all the processes explained in this report we created a process diagram describing each major step in generating profile-based news recommendations.

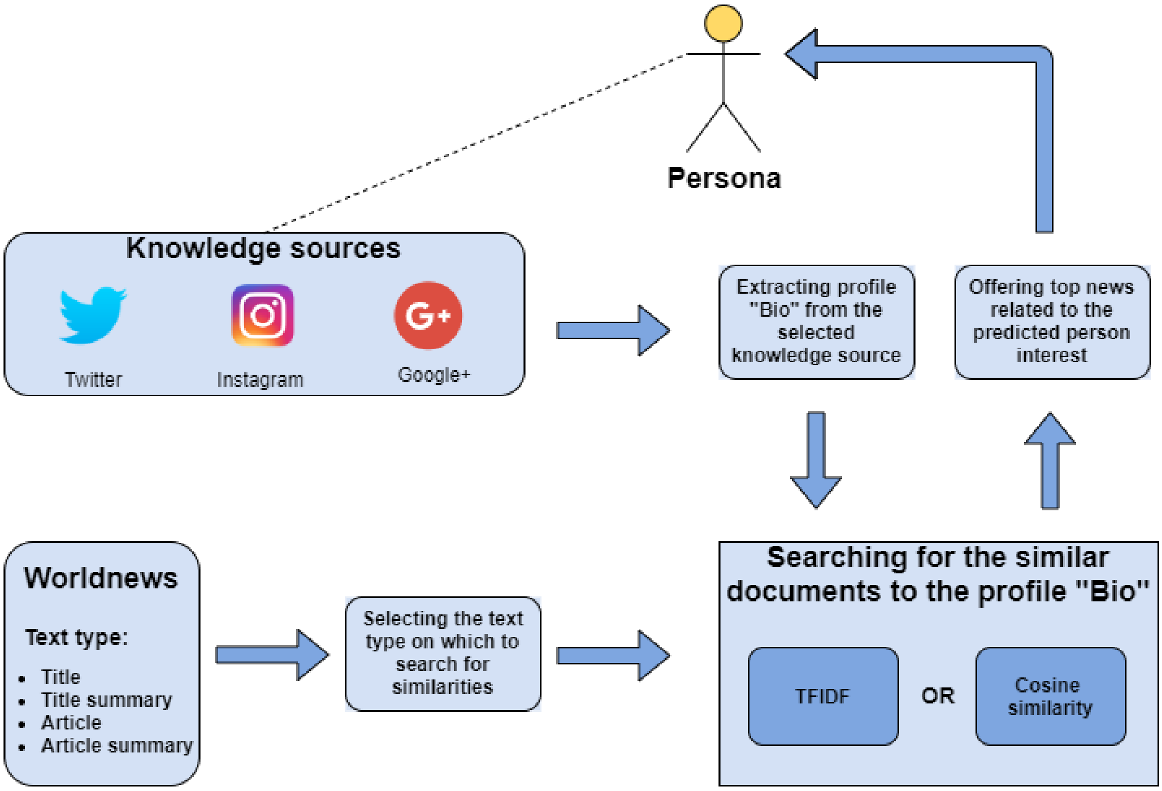

Persona
Knowledge sources
Twitter
Instagram
Google+
Extracting profile "Bio" from the selected knowledge source
Offering top news related to the predicted person interest
Worldnews
Text type:
Title
Title summary
Article
Article summary
Selecting the text type on which to search for similarities
Searching for the similar documents to the profile "Bio"
TFIDF
OR
Cosine similarity

Figure 5: Overview of the profile-based news recommendation

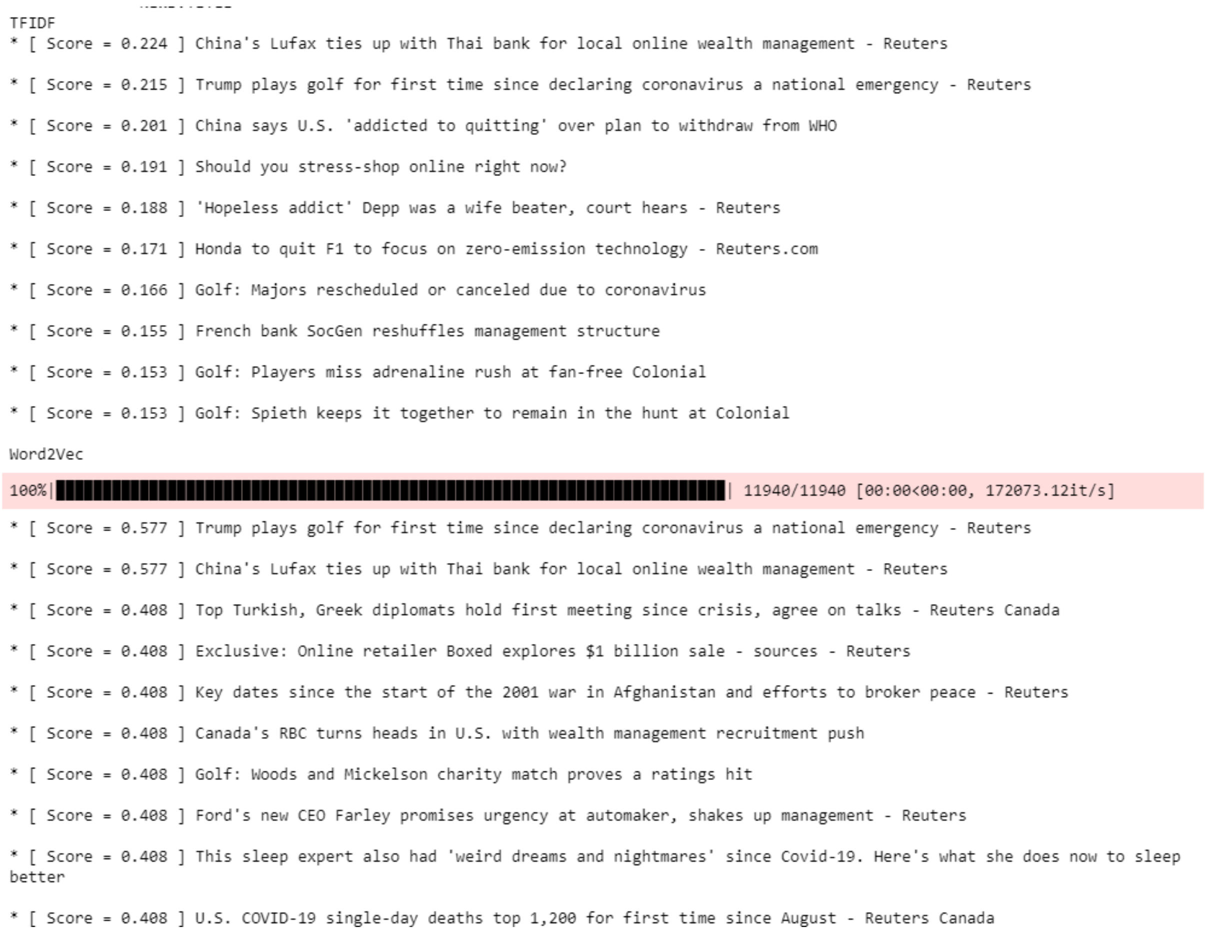
Now, let's have a look at the a practical example we obtained from selecting a random person and applying the profile information we obtained from Twitter. This person is: *Francesco Venier*. We took his twitter Bio: "prof unitrieste mibschool curious technology management bmwgssupslowfoodtravel golf photography addict online since"

In the images below one could find 4 settings:

- News title + TFIDF
- News title + Word2Vec
- News description + TFIDF
- News description + Word2Vec

The firs two results represent the usage of news title. Hense, we have few information to understand where the news article could be recommended. As we can see in 4 there are similar news recommendations produced by both approaches. But if we take a closer look TFIDF took into account a word "golf" that has a high importance score that resulted in receiving more results related to "golf". Overall we could see that most of the news except from golf are related to "technology", "management" and "online" mentioned in the person's Bio.

On the other hand, the Word2Vec approach produced the recommendation that address mostly politics and economics in the world. One could infer that is is a result address the overall combination of words and their distances between each other creating a a robust context. Also it is noticeable, that ”golf” here is recommended 2 times only and in the context where it is not key aspect of the news: ” Trump plays golf for first time since declaring coronavirus a national emergency - Reuters” is about emergency caused by coronavirus and ”Golf: Woods and Mickelson charity match proves a ratings hit” central point here is charity that draws people attention to watching golf.

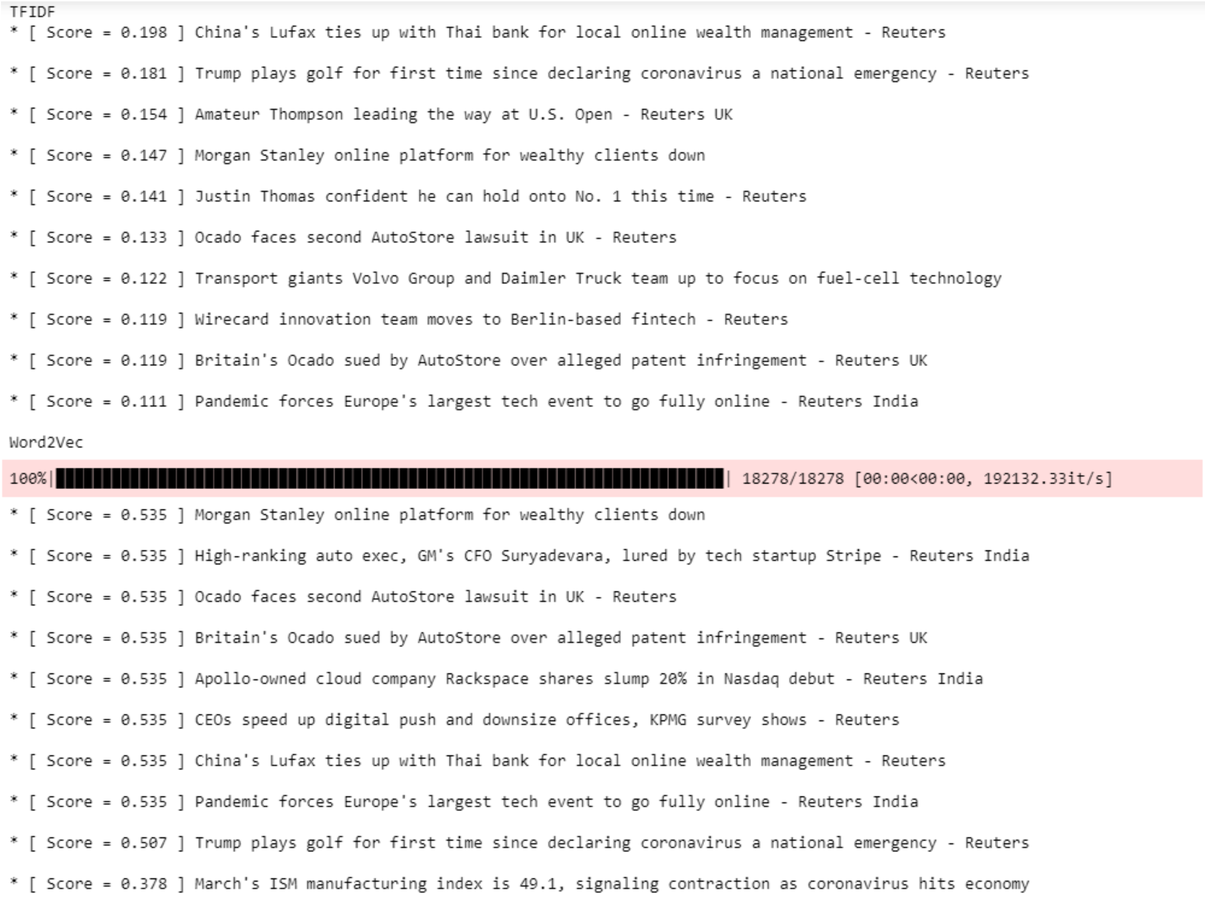


```
TFIDF
* [ Score = 0.224 ] China's Lufax ties up with Thai bank for local online wealth management - Reuters

* [ Score = 0.215 ] Trump plays golf for first time since declaring coronavirus a national emergency - Reuters

* [ Score = 0.201 ] China says U.S. 'addicted to quitting' over plan to withdraw from WHO

* [ Score = 0.191 ] Should you stress-shop online right now?

* [ Score = 0.188 ] 'Hopeless addict' Depp was a wife beater, court hears - Reuters

* [ Score = 0.171 ] Honda to quit F1 to focus on zero-emission technology - Reuters.com

* [ Score = 0.166 ] Golf: Majors rescheduled or canceled due to coronavirus

* [ Score = 0.155 ] French bank SocGen reshuffles management structure

* [ Score = 0.153 ] Golf: Players miss adrenaline rush at fan-free Colonial

* [ Score = 0.153 ] Golf: Spieth keeps it together to remain in the hunt at Colonial

Word2Vec
100%|██████████████████████████████████████████████████████████████████████| 11940/11940 [00:00<00:00, 172073.12it/s]
* [ Score = 0.577 ] Trump plays golf for first time since declaring coronavirus a national emergency - Reuters

* [ Score = 0.577 ] China's Lufax ties up with Thai bank for local online wealth management - Reuters

* [ Score = 0.408 ] Top Turkish, Greek diplomats hold first meeting since crisis, agree on talks - Reuters Canada

* [ Score = 0.408 ] Exclusive: Online retailer Boxed explores $1 billion sale - sources - Reuters

* [ Score = 0.408 ] Key dates since the start of the 2001 war in Afghanistan and efforts to broker peace - Reuters

* [ Score = 0.408 ] Canada's RBC turns heads in U.S. with wealth management recruitment push

* [ Score = 0.408 ] Golf: Woods and Mickelson charity match proves a ratings hit

* [ Score = 0.408 ] Ford's new CEO Farley promises urgency at automaker, shakes up management - Reuters

* [ Score = 0.408 ] This sleep expert also had 'weird dreams and nightmares' since Covid-19. Here's what she does now to sleep
better

* [ Score = 0.408 ] U.S. COVID-19 single-day deaths top 1,200 for first time since August - Reuters Canada
```

Figure 6: TFIDF and Word2Vec for news TITLE

The previous examples judged whether to recommend the news based only on the article title. In the next two approaches 4 we will use the news description as a broader source of information. Here we hope that more information will help us address interests better than just a title.

We could find out that the result of the TFIDF and Word2Vec approaches improves. As now ”golf” doesn’t play such important role and the news are now generated giving more value to other words mentioned in the profile description. Talking about Word2Vec one could notice that the news recommendation moved closely to business topics. We presume that overall combination of words in person’s Bio have short distances to business related words in Word2Vec which could explain such recommendation results.



```
TFIDF
* [ Score = 0.198 ] China's Lufax ties up with Thai bank for local online wealth management - Reuters

* [ Score = 0.181 ] Trump plays golf for first time since declaring coronavirus a national emergency - Reuters

* [ Score = 0.154 ] Amateur Thompson leading the way at U.S. Open - Reuters UK

* [ Score = 0.147 ] Morgan Stanley online platform for wealthy clients down

* [ Score = 0.141 ] Justin Thomas confident he can hold onto No. 1 this time - Reuters

* [ Score = 0.133 ] Ocado faces second AutoStore lawsuit in UK - Reuters

* [ Score = 0.122 ] Transport giants Volvo Group and Daimler Truck team up to focus on fuel-cell technology

* [ Score = 0.119 ] Wirecard innovation team moves to Berlin-based fintech - Reuters

* [ Score = 0.119 ] Britain's Ocado sued by AutoStore over alleged patent infringement - Reuters UK

* [ Score = 0.111 ] Pandemic forces Europe's largest tech event to go fully online - Reuters India

Word2Vec
100%|██████████████████████████████████████████████████████████████████████| 18278/18278 [00:00<00:00, 192132.33it/s]
* [ Score = 0.535 ] Morgan Stanley online platform for wealthy clients down

* [ Score = 0.535 ] High-ranking auto exec, GM's CFO Suryadevara, lured by tech startup Stripe - Reuters India

* [ Score = 0.535 ] Ocado faces second AutoStore lawsuit in UK - Reuters

* [ Score = 0.535 ] Britain's Ocado sued by AutoStore over alleged patent infringement - Reuters UK

* [ Score = 0.535 ] Apollo-owned cloud company Rackspace shares slump 20% in Nasdaq debut - Reuters India

* [ Score = 0.535 ] CEOs speed up digital push and downsize offices, KPMG survey shows - Reuters

* [ Score = 0.535 ] China's Lufax ties up with Thai bank for local online wealth management - Reuters

* [ Score = 0.535 ] Pandemic forces Europe's largest tech event to go fully online - Reuters India

* [ Score = 0.507 ] Trump plays golf for first time since declaring coronavirus a national emergency - Reuters

* [ Score = 0.378 ] March's ISM manufacturing index is 49.1, signaling contraction as coronavirus hits economy
```

Figure 7: TFIDF and Word2Vec for news DESCRIPTION

5 Conclusion

In this report we presented a solution for the news recommendation system that takes social media profile information and finds the most similar news articles to offer to a respective person - social network account holder.

To sum up, we believe that both approaches of recommender systems are applicable for the real news recommendation. Overall, we presume taking the context of words and distances to the other words as more intriguing approach for such problems and if we happen to select between TFIDF and Word2Vec we would move with the last one.

In addition, one of the important insights in this report is the approach of taking profile descriptions from social network could supply with enough information to provide the news recommendations.