NLP Task

SkillMapper

Issues to solve

- Commodification of courses
  - **Too many courses** to pick from
- Limited exposure to reviews
  - **Too many reviews**
  - Good reviews always on top
- High search time
  - Comparing multiple courses
  - Caused by the 1st issue

Issues to solve

- Commodification of courses
  - **Too many courses** to pick from

Something that takes into account **all the courses**

- Limited exposure to reviews
  - **Too many reviews**
  - Good reviews always on top

Something that takes into account **all the reviews**

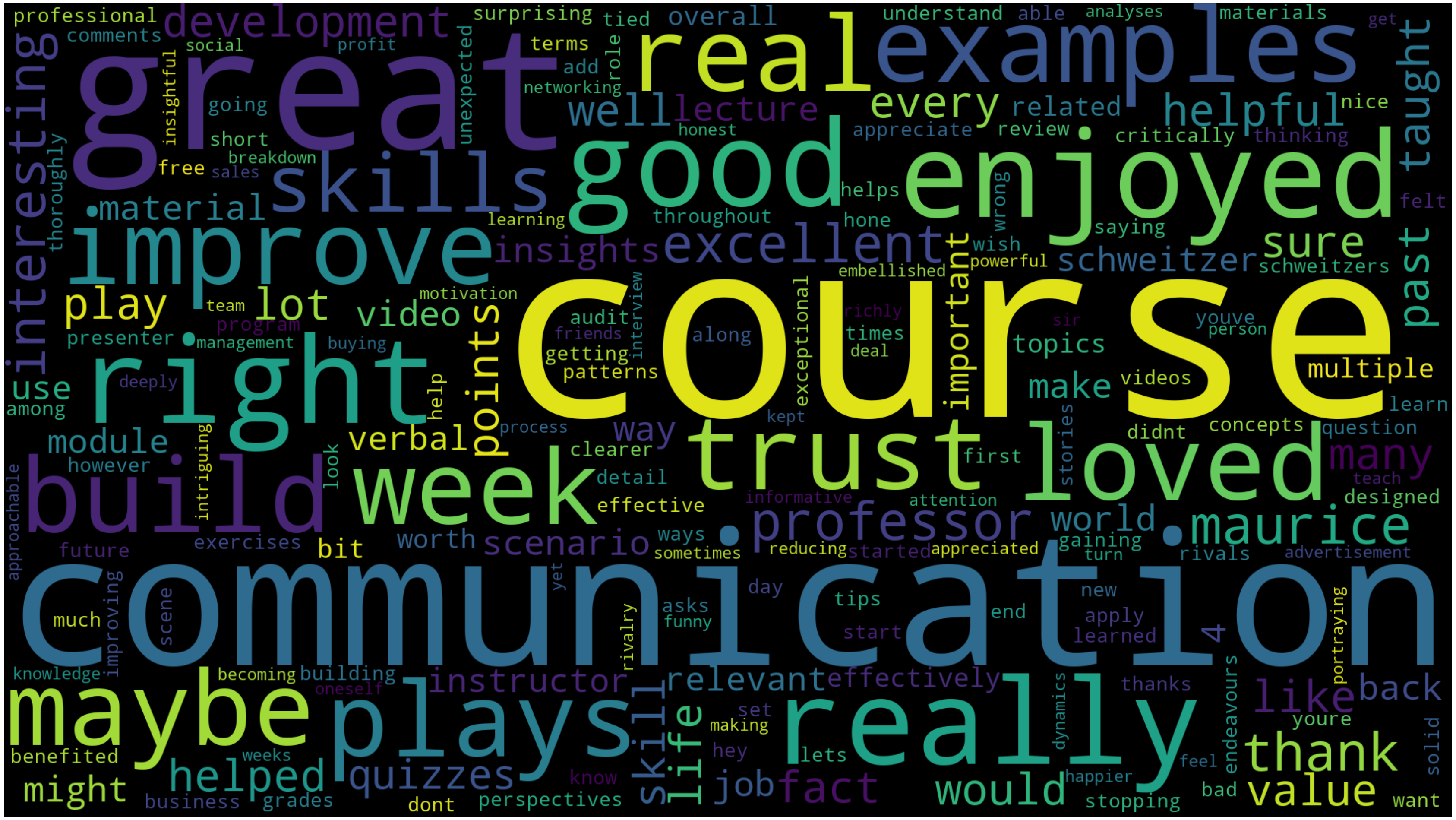
- High search time
  - Comparing multiple courses
  - Caused by the 1st issue

Decision making should not take away too much time

Possible Approaches

- Wordcloud and top-words
- Sentiment Analysis
- TF-IDF Approach
- Cosine similarity using spaCy
- Extra metadata

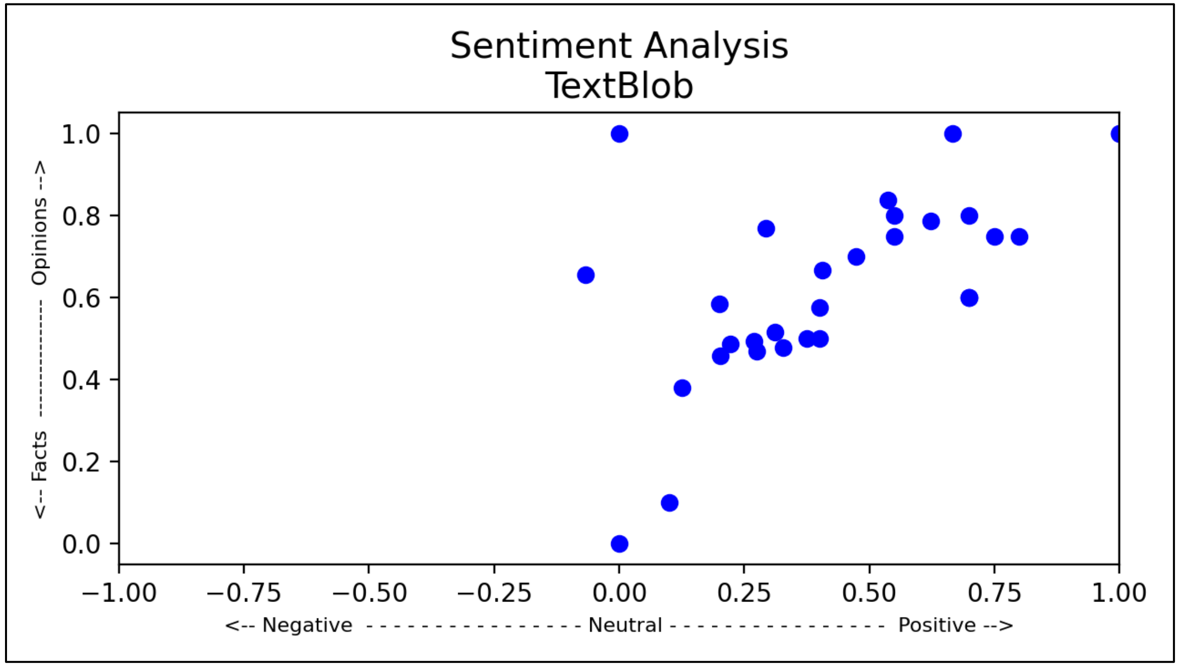
1. Wordcloud & Top-Words

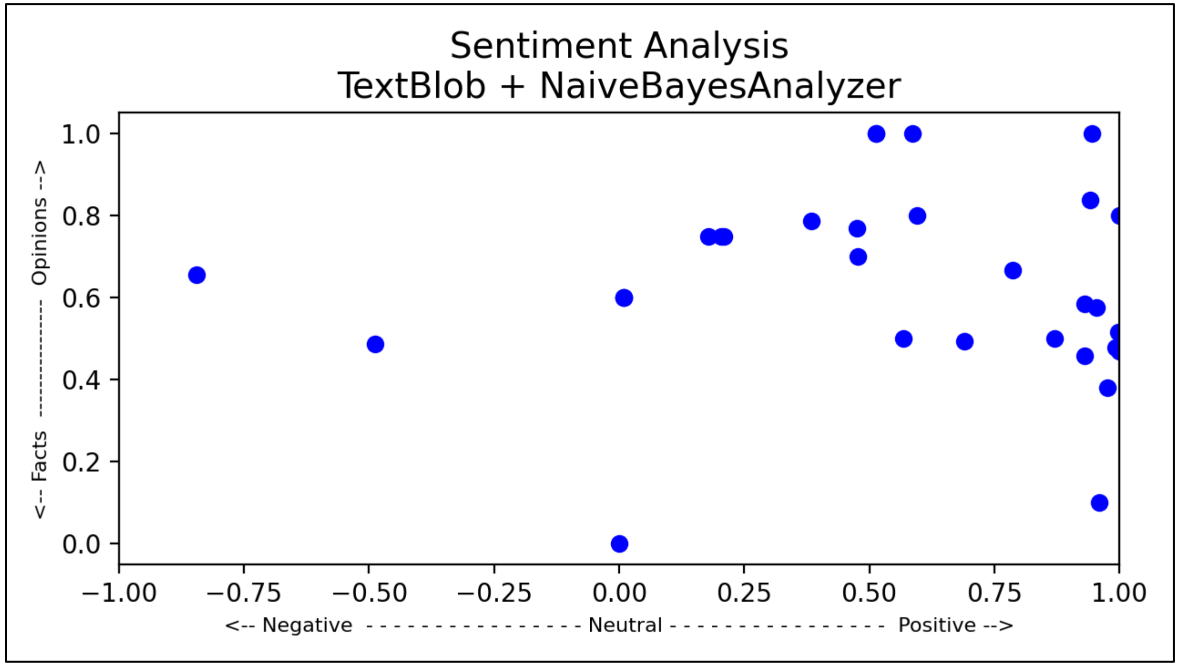


word	count
course	23
communication	12
great	10
enjoyed	4
right	4
good	4
really	4
plays	3
improve	3
loved	3
real	3
examples	3
week	3
maybe	3
build	3
trust	3
skills	3
professor	3
thank	3
maurice	3
interesting	3

2. Sentiment Analysis

- TextBlob (based on NLTK)
  - PatternAnalyzer
  - NaiveBayesAnalyzer
- Possible for each course
  - Comparison between courses





3. TF-IDF Approach

- Generally used for words with corpus (document)
- Can be *possibly* leveraged for reviews

TF-IDF Approach

'Gold-standard'

Course α

review 1
review 2
review 3
.
.
.
review n

Course 1

review 1
review 2
review 3
.
.
.
review n

Course 2

review 1
review 2
review 3
.
.
.
review n

. . .

Course M

review 1
review 2
review 3
.
.
.
review n

TF-IDF Approach

$$TF * IDF$$
$$= Term\ Frequency\ *\ Inverse\ Document\ Frequency$$

$$= \frac{Occurences\ of\ a\ word[i]\ in\ course[j]'s\ review}{total\ unique\ words\ in\ course[j]} * \log\left(\frac{Total\ no.of\ courses}{Occurences\ of\ word[i]across\ courses}\right)$$

Course 1	Course 2	. . .	Course M
review 1	review 1		review 1
review 2	review 2		review 2
review 3	review 3		review 3
.	.		.
.	.		.
.	.		.
review n	review n		review n

word	count
course	23
communication	12
great	10
enjoyed	4
right	4
good	4
really	4
plays	3
improve	3
loved	3
real	3
examples	3
week	3
maybe	3
build	3
trust	3
skills	3
professor	3
thank	3
maurice	3
interesting	3

TF-IDF Approach

$$\begin{array}{cc}
 & \begin{matrix} W1 & W2 & \dots & W_N \end{matrix} \\
\begin{matrix} \text{Course 1} \\ \text{Course 2} \\ \cdot \\ \cdot \\ \cdot \\ \text{Course M} \end{matrix} & \begin{bmatrix} \quad & \quad & \quad & \quad \\ & & & \\ & & & \\ & & & \\ & & & \\ & & & \end{bmatrix}
\end{array}$$

$$TF * IDF$$
$$= Term\ Frequency\ *\ Inverse\ Document\ Frequency$$
$$= \frac{Occurences\ of\ a\ word[i]\ in\ course[j]'s\ review}{total\ unique\ words\ in\ course[j]} * log\left(\frac{Total\ no.\ of\ courses}{Occurences\ of\ word[i] across\ courses}\right)$$

Takes into account all reviews

Takes into account all courses

TF-IDF Approach

$$\begin{array}{cc} & \begin{matrix} W1 & W2 & \dots & W_N \end{matrix} \\ \begin{matrix}\text{Course 1}\\ \text{Course 2}\\ .\\ .\\ .\\ \text{Course M}\end{matrix} & \begin{bmatrix} \quad & \quad & \quad & \quad \\ \\ \\ \\ \\ \\ \end{bmatrix}\end{array} = \begin{bmatrix}\text{vector 1}\\ \text{vector 2}\\ .\\ .\\ .\\ \text{vector M}\end{bmatrix}$$

4. spaCy's `.similarity()`

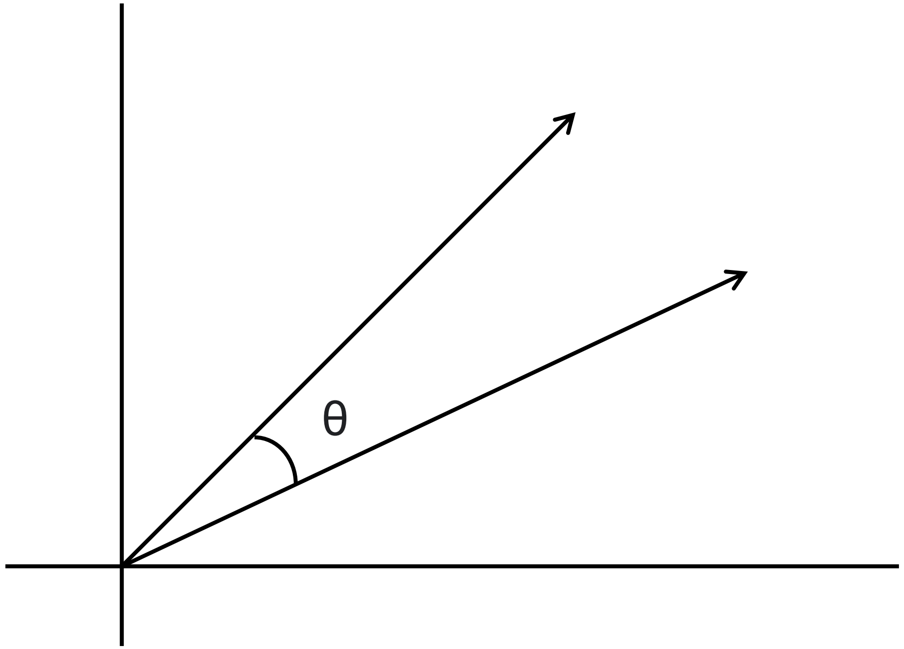
- Based on cosine similarity
- Returns value between [0, 1]

```
nlp = spacy.load("en_core_web_sm")

doc_gold = nlp(course_gold_reviews)

doc1 = nlp(course1_reviews)
doc2 = nlp(course2_reviews)
doc3 = nlp(course3_reviews)

doc_gold.similarity(doc1)
doc_gold.similarity(doc2)
doc_gold.similarity(doc3)
```



- `vector` computed using word2vec technique

```
return numpy.dot(self.vector, other.vector) / (self.vector_norm * other.vector_norm)
```

5. Extra Metadata

```
id_review; id_course; url; rating; platform; review; language
```

coursera Explore

What do you want to learn?

For Enterprise For Students Archit Yadav

Filter by: All Learners All Stars

Sort by: **Most Helpful**

1 - 25 of 68 Reviews for Introduction to Machine Learning in Production

By Francisco R • May 21, 2021

I know it's an introduction, but I got a bit disappointed. It's quite basic and even though it has some hands on notebooks, they're optional and you don't need to work on anything. Quizzes are easy, and I didn't have the feeling I learnt much. I'm still rating it with 3 because, well, it's Andrew Ng, and this his teaching is worth gold.

This is helpful (3)

By Mohamed A H • May 14, 2021

I give you the full review stars since I learned many new things that I did not pay attention to before, e.g.: I used to focus on models for many years instead of data.

This is helpful (2)

By HARI A K • May 16, 2021

Really good for anyone with strong background in DL and ML... And want to be able to start a real time project... Or lead a ML team

This is helpful (2)

By Wesley E B • May 16, 2021

It had some great advice for how to design a machine learning system. More practical examples would have been appreciated.

This is helpful (1)

5. Extra Metadata

word	count
course	23
communication	12
great	10
enjoyed	4
right	4
good	4
really	4
plays	3
improve	3
loved	3
real	3
examples	3
week	3
maybe	3
build	3
trust	3
skills	3
professor	3
thank	3
maurice	3
interesting	3

$$\begin{bmatrix} 3 \\ 2 \\ 2 \\ 1 \end{bmatrix}$$

Course 1
Course 2
.
.
.
Course M

$W1 \quad W2 \quad \ldots \quad W_N$

$$TF * IDF$$
$$= Term\ Frequency\ *\ Inverse\ Document\ Frequency$$

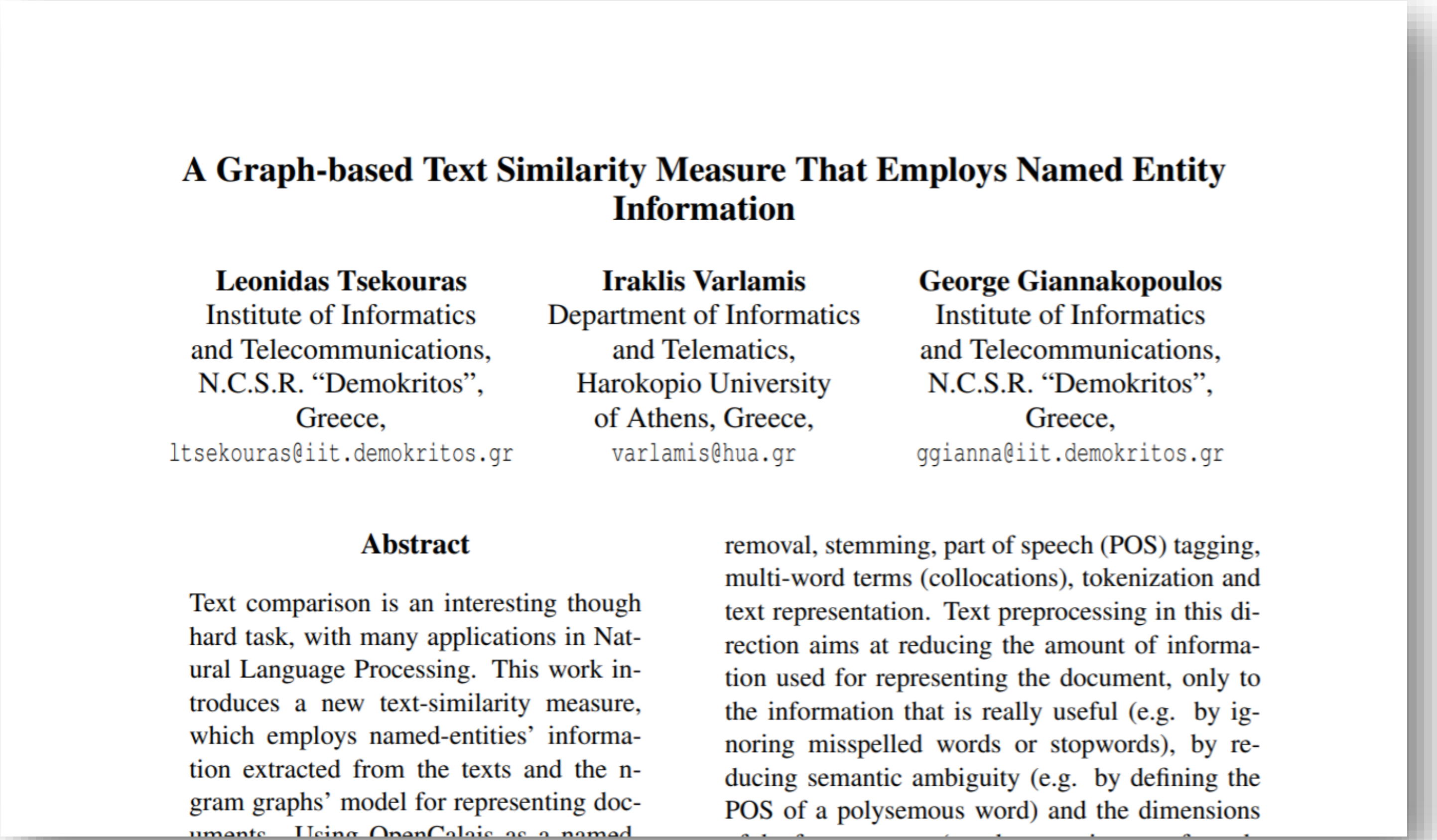
$$= \frac{Occurences\ of\ a\ word[i]\ in\ course[j]'s\ review}{total\ unique\ words\ in\ course[j]} * \log\left(\frac{Total\ no.of\ courses}{Occurences\ of\ word[i] across\ courses}\right) * cmt\_weight$$

Possible Approaches

1. Wordcloud and top-words
2. Sentiment Analysis
3. TF-IDF Approach
4. Cosine similarity using spaCy
5. Extra metadata

Combination of some/all?

Named Entity Recognition (NER)



A Graph-based Text Similarity Measure That Employs Named Entity Information

Leonidas Tsekouras
Institute of Informatics
and Telecommunications,
N.C.S.R. "Demokritos",
Greece,
ltsekouras@iit.demokritos.gr

Iraklis Varlamis
Department of Informatics
and Telematics,
Harokopio University
of Athens, Greece,
varlamis@hua.gr

George Giannakopoulos
Institute of Informatics
and Telecommunications,
N.C.S.R. "Demokritos",
Greece,
ggianna@iit.demokritos.gr

Abstract

Text comparison is an interesting though hard task, with many applications in Natural Language Processing. This work introduces a new text-similarity measure, which employs named-entities' information extracted from the texts and the n-gram graphs' model for representing documents.

removal, stemming, part of speech (POS) tagging, multi-word terms (collocations), tokenization and text representation. Text preprocessing in this direction aims at reducing the amount of information used for representing the document, only to the information that is really useful (e.g. by ignoring misspelled words or stopwords), by reducing semantic ambiguity (e.g. by defining the POS of a polysemous word) and the dimensions