

Natural Language Processing

Text Similarity in Essays

Alexander Garcia

October 2024

Contents

1	Introduction	1
2	Shape of the Data	2
3	Method 1: Word frequency-based approach	2
3.1	Example of Jaccard	2
3.2	Limitations	3
3.3	Results of Jaccard	3
4	TF-IDF	3
5	Method 2: Cosine Similarity	4
5.1	Results of Cosine Similarity	5
6	Method 3: TF-IDF Euclidean Distance	6
6.1	Results of Euclidean Distance	6
7	Conclusion	6

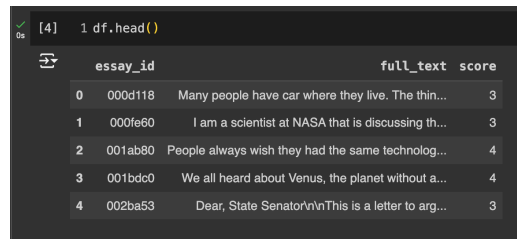
1 Introduction

The goal of this project is to utilize different methods for finding similarities between essays in the Automated Essay Scoring 2.0 dataset. The first method implemented is Jaccard Similarity which bases similarity on the intersection

and union of words between two documents. The other 2 methods, Cosine Similarity and Euclidean Distance, are based on text vectorization which results in improvements to both speed and accuracy of determining which documents are most similar.

2 Shape of the Data

The csv file contains 17,307 essays. As seen in the figure, each row contains an essay ID, the text of the essay, and a holistic score of the essay. However, note that score is not used in any of the methods to determine word-frequency similarity.



```
[4] 1 df.head()
```

	essay_id	full_text	score
0	000d118	Many people have car where they live. The thin...	3
1	000fe60	I am a scientist at NASA that is discussing th...	3
2	001ab80	People always wish they had the same technolog...	4
3	001bdc0	We all heard about Venus, the planet without a...	4
4	002ba53	Dear, State Senator\r\nThis is a letter to arg...	3

Figure 1: Head of the data frame

3 Method 1: Word frequency-based approach

At first the data is looped through to clean some symbols and remove stop words like “of”, “a”, “the”, etc... using the nltk library. For the first method, I implemented a modified Jaccard Similarity so it would include the frequency of words instead of simply the appearance of one word. This works by taking the sum of the frequencies of intersecting words and divides that result by the sum of the frequencies of the union of words between two documents.

$$\text{Jaccard Similarity} = \frac{\text{Sum of Intersection Frequencies}}{\text{Sum of Union Frequencies}}$$

3.1 Example of Jaccard

Essay 1: {car: 1, blah: 1, test: 1}

Essay 2: {car: 1, bingo: 1, test: 1}

Intersection = 1(car) + 1(test) = 2

Union = 1(car) + 1(blah) + 1(bingo) + 1(test) = 4

Jaccard Similarity = $\frac{2}{4} = 0.5$

3.2 Limitations

Time complexity is the biggest limitation of this algorithm. One loop was needed to control the baseline essay and another inner loop to iterate through the remaining essays to be compared against, such that essay[i] would be compared with essay[i+1] through N (17307). Within those two loops there is also a need to iterate through the dictionary of word frequencies for both essays to get their sums. This results in a running time of approximately $O(n^4)$ making this algorithm impractical on large datasets

3.3 Results of Jaccard

While unable to run this algorithm on all 17k essays, I did narrow the loop iteration down such that the first 500 essays in the dataset are compared with the following 10,000.

```
Essay 1: 0792285
Essay 1 text: Dear State Senator,
I think that the Electoral College is important because it is a process. The founding fathers estab
Essay 2: 4afbbfa
Essay 2 text: Dear, To whom ever it may concern I here am writing this this letter to tell you guys about the Electoral College or cha
Jaccard similarity score: 0.4615

Essay 1: 07a14a5
Essay 1 text: The author's claim of studying Venus is a worthy pursuit because Venus is closely related to Earth, Venus has a envirome
Essay 2: 8fa257f
Essay 2 text: "The Challengen of Exploring Venus," The author suggests that studying Venus is a worthy pursuit despite the dangers it p
Jaccard similarity score: 0.4272
```

Figure 2: Excerpts of two essays and their Jaccard similarity score

As seen in the figure for the first two sets of essays, they are both letters addressed to political figures discussing the Electoral College. The second set of essays have a similar topic of studying the planet Venus. Even though the scores have a low score they ranked among the highest 5 relative to all the essays compared.

4 TF-IDF

TF-IDF stands for **term frequency-inverse document frequency**. This strategy of vectorization turns the text into a numerical representation based

(0, 27742)	0.01020199990972020499
(0, 34386)	0.013303624459340689
(0, 26538)	0.016029338203867188
(0, 30309)	0.1544227362835433
(0, 976)	0.0910318292040356
(0, 40284)	0.024538905016586097
:	:
(17306, 25154)	0.09194363808557424
(17306, 63054)	0.18078423429068585
(17306, 59807)	0.07101848485017499
(17306, 26561)	0.063194167929799
(17306, 15003)	0.06821118030427632
(17306, 19827)	0.0839222198970375
(17306, 35928)	0.11521701127181959
(17306, 49525)	0.1281074196254864

Figure 3: TF-IDF output

on word frequency relative to the entire corpus of text. The tuple represents the document index and word index respectively i.e., (document index, word index). The value represents the TF-IDF score of a words importance within the corpus.

5 Method 2: Cosine Similarity

This method is based on TF-IDF vectorization which has significant improvements to the running time. Compared to the Jaccard, which didn't finish, vectorization on the entire corpus only took a few seconds and Cosine Similarity took between 30-40 seconds on each run. Cosine similarity works by taking the cosine of the angle between two vectors. The closer together two vectors are, the smaller the angle, which results in a higher cosine. The cosine similarity between vectors is represented by the following formula

$$\text{similarity}(A,B) = \frac{A \cdot B}{||A|| ||B||}$$

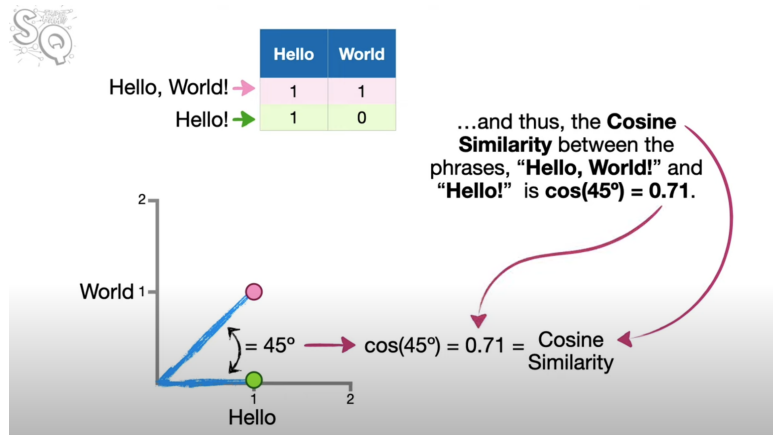


Figure 4: Cosine similarity [Sta]

5.1 Results of Cosine Similarity

The only stop word removed in this method was "PROPER_NAME" as there appeared to be duplicate essays in which "PROPER_NAME" was present. The algorithm seems to be accurate as it correctly detects the essays shown below as being similar with a similarity score of 0.80.

```

1 df.iloc[2708]["full_text"]
--NORMAL--
'I think the Electoral college is a good way to vote for the president or vice president because i think without the Electoral college, we wouldn't have anyone to vote for the president or vice president. We have each candidate running for president in our state and it has his or her own group of electors. The electors are chosen by the candidate political party.\n\nI agree that most states have a winner take all system because it awards all electors to the winning presidential candidate. The 23rd Amendment of the constitution, the District of Columbia is allocated 3 electors and treated like a state for purposes of the Electoral College.\n\nThe word state refers to the District of Columbia. The Electoral College is a process, not a place. The founding fathers established it in the Constitution as a compromise between election of the President by a vote in congress and election of the President by a popular vote of qualified citizens. We help choose our state's electors when we vote for ...'

1 df.iloc[6132]["full_text"]
'I think that we should keep the Electoral College but we could also change election by popular vote because there could be problems over the outcome of an Electoral College, the reason is that the winning candidate's share of the Electoral College invariably exceeds his share of the popular vote. The Electoral College requires a presidential candidate to have trans-regional appeal. A candidate with only regional appeal is unlikely to be a successful president. Voters in toss-up are more likely to pay close attention to the campaign, knowing that they are going to decide the election. The Electoral College avoids the problem of elections in which no candidate receives a majority of the vote cast. It also restores some weight in the political balance. There is pressure for run-off elections when no candidate wins a majority of the votes cast. The Electoral College method of selecting the president may turn off potential voters for a candidate who has no hope of carrying their state.\n\n...'

```

Figure 5: Excerpts of two essays deemed similar using Cosine-Similarity

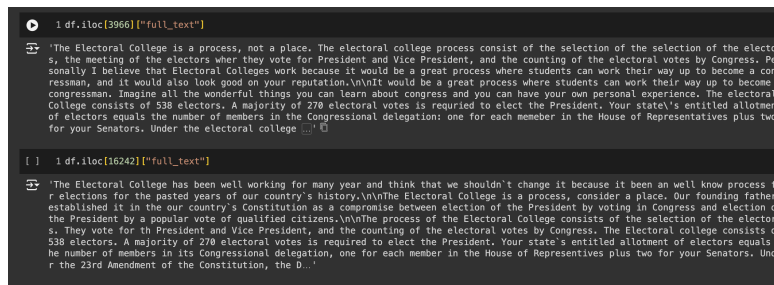
As seen in the figure, these two essays are discussing the similar topic of the Electoral College.

6 Method 3: TF-IDF Euclidean Distance

Similar to Cosine Similarity, the Euclidean Distance method also utilizes vectors. The Euclidean approach is slightly simpler to understand as it measures the distance between two points on vectors instead of using the angle between them. A lower result from Euclidean means the vectors are closer together and thus more similar than vectors which have a large value.

6.1 Results of Euclidean Distance

The Euclidean method results in similar accuracy as the Cosine Similarity. I narrowed the top essays down to a search of essays which had a distance less than 0.5 (meaning they would be similar) but above 0 to filter out an essay being matched with itself.



```
1 df.iloc[3966]["full_text"]
"The Electoral College is a process, not a place. The electoral college process consist of the selection of the elector
s, the meeting of the electors wher they vote for President and Vice President, and the counting of the electoral votes by Congress. Per
sonally I believe that Electoral Colleges work because it would be a great process where students can work their way up to become a cong
ressman, and it would also look good on your reputation.\n\nIt would be a great process where students can work their way up to become a
congressman. Imagine all the wonderful things you can learn about congress and you can have your own personal experience. The electoral
College consists of 538 electors. A majority of 270 electoral votes is required to elect the President. Your state's entitled allotment
of electors equals the number of members in the Congressional delegation: one for each member in the House of Representatives plus two
for your Senators. Under the electoral college ...

1 df.iloc[16242]["full_text"]
"The Electoral College has been well working for many year and think that we shouldn't change it because it been an well know process fo
r elections for the pasted years of our country's history.\n\nThe Electoral College is a process, consider a place. Our founding fathers
established it in the our country's Constitution as a compromise between election of the President by voting in Congress and election of
the President by a popular vote of qualified citizens.\n\nThe process of the Electoral College consists of the selection of the elector
s. They vote for th President and Vice President, and the counting of the electoral votes by Congress. The Electoral college consists of
538 electors. A majority of 270 electoral votes is required to elect the President. Your state's entitled allotment of electors equals t
he number of members in its Congressional delegation, one for each member in the House of Representatives plus two for your Senators. Unde
r the 23rd Amendment of the Constitution, the D...
```

Figure 6: Excerpts of two essays deemed similar using Euclidean Distance

As seen in the figure above, the algorithm correctly picks essays that are both discussing the Electoral College. Note, these essays picked are different Electoral College essays than Cosine Similarity reported.

7 Conclusion

The results of the Jaccard similarity seem to accurately find essays that are similar but the time complexity of that method makes it impractical on large corpus with many documents. Vectorization shows massive improvements to both time and accuracy as seen in the Cosine Similarity and Euclidean Distance methods. All three methods correctly identified essays which were not similar in topic by assigning them low similarity scores.

References

- [Sta] Josh Starmer. *Cosine Similarity, Clearly Explained!!!* URL: <https://www.youtube.com/watch?v=e9U0QAFbfLI>.