Polyratings TF IDF

Max Fisher, Kyle Hultgren, Nick Allen

Goal

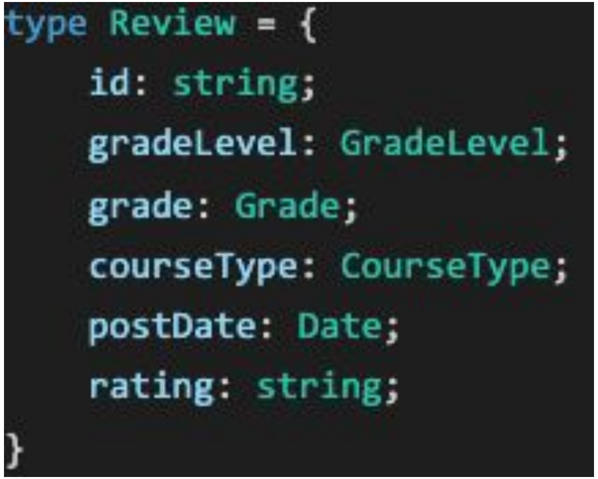
- Search through polyratings reviews using tf idf and cosine similarity
- Can search using long text input and searches by looking for most “impactful” words

$$w_{i,j} = tf_{i,j} \times log(\frac{N}{df_i})$$

$$\frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}},$$

Data

- Json file format
- From Polyratings website
- 48,000+ reviews



```
type Review = {
    id: string;
    gradeLevel: GradeLevel;
    grade: Grade;
    courseType: CourseType;
    postDate: Date;
    rating: string;
}
```

Term Frequency

- How many times a term appears in a document
- Normalized term frequency gives a more accurate weight of term
- Term Freq. / total words

Inverse Document Frequency

- Document frequency is the amount of times a word appears across all documents
    - $\{d \in D \mid t \in d\}$ where: D is the document corpus, and t is the target term
- Inverse document frequency is just the log of the number of documents divided by the document frequency.
- More common words across the whole document corpus will result in a significantly lower inverse document frequency
- TF IDF is just the term frequency multiplied by the inverse document frequency to get a measure of how “important” a word is in a given document

Query TF IDF

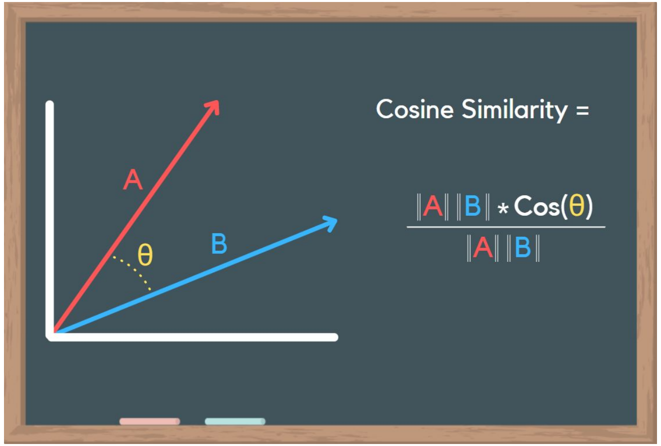
1. Break up the query into words
2. Calculate the term frequency for every unique word
3. Do an inner join on the query to the calculated idf for the document copus
4. Multiply the term frequency to the idf to get the query tf idf

Result is a Data Frame that maps query words to their importance

Query Word	Tf idf
wonderful	0.54
and	0.03

Cosine Similarity

- Used to compare similarity between query and document
- Higher number means higher similarity
- The closer the two vectors are means a smaller angle between them. Smaller angle = Larger Cosine
- Larger Cosine = Larger Similarity



$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

Demo