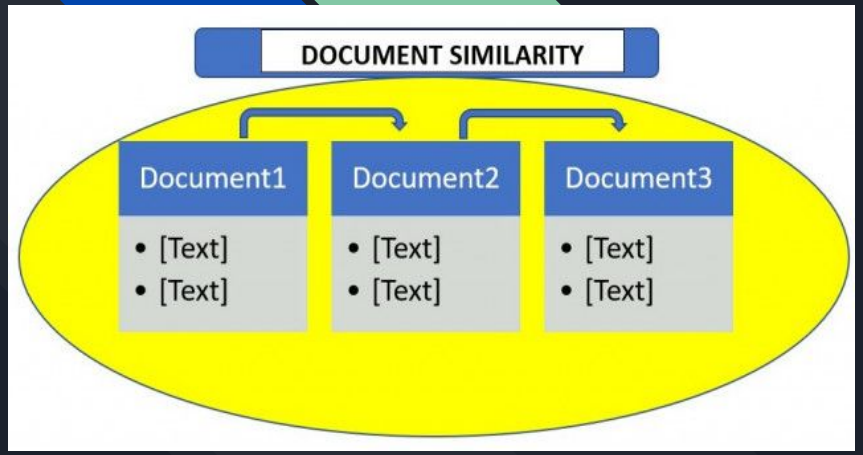
DOCUMENT SIMILARITY



Team members:

Siri Koduru
Lakshmi Sahithi Yalamarthy

Problem statement

- How to detect similarity between documents among huge number of documents.
- Main aim is to find text similarities with text embeddings
- Represent text into vector space.
- Major aspects
  - negative sampling
  - Four-fold classification

Motivation

- Aspect Based Document Similarity for Research Papers
- Plagiarism detection

Proposed approach

- Input Dataset - 20 new group dataset
- Split input data set into train and test sets
- Preprocessed the data - tagging, removal of stop words and punctuations
- Algorithms - TFIDF, word2vec, doc2vec, BERT, and glove
- Cosine and euclidean distances on each algorithm
- Baseline Methods : Validation-cross validation on categories, avg-vector similarity(mean similarity)

Results

All the results are attached below

The documents that are identified similar all speak about selling a product that is the similarity identified from the data set of news papers

Results cont..

```
From: dtmedin@catbyte.b30.ingr.com (Dave Medin)
Subject: Pressure meter
Reply-To: dtmedin@catbyte.b30.ingr.com
Organization: Intergraph Corporation, Huntsville AL
Lines: 21

Heise model 710A pressure meter. This is a precision 4-1/2
digit meter measuring 0 - 15 PSI (absolute) in .001 psi
increments.  Case is in extremely good shape, and can be used
as a stand-alone meter or panel mounted. Brass fitting (looks
like standard 3/8") on back. Operates from 110 VAC.

I'd like $50 for it, or make an offer. It is a lot more useful to
a lab than as an ersatz barometer, which is what I've been using
it for.

--
-------------------------------------------------------------------
       Dave Medin                          Phone:   (205) 730-3169 (w)
    SSD--Networking                                 (205) 837-1174 (h)
    Intergraph Corp.
       M/S GD3004                  Internet: dtmedin@catbyte.b30.ingr.com
  Huntsville, AL 35894             UUCP:  ...uunet!ingr!b30!catbyte!dtmedin

   ******* Everywhere You Look (at least around my office) *******

 * The opinions expressed here are mine (or those of my machine)
```

From: HO@kcgl1.eng.ohio-state.edu (Francis Ho)
Subject: 286 Laptop
Nntp-Posting-Host: kcgl1.eng.ohio-state.edu
Organization: The Ohio State University
Lines: 18

MITSBISHI Laptop (MP 286L)

-286/12 (12,8,6 MHz switchable)
-2M RAM installed
-Backlit CGA (Ext. CGA, MGA)
-20M 3.5"HH HDD/1.44M 3.5" FDD
-2 COM/1 LPT ports
-complete manual set
-Built like a tank
-Excellent cosmetic cond.
-dark gray
-used very lightly

Problems:
(1)HDD stops working.
(2)LCD sometimes doesn't work (ext. CAG/MGA works).

Best Offer.

Mean and average similarity in 4 groups of news data in cross validation

```
Category: sci.electronics
    Mean difference: 0.11, Same-category average similarity: 0.3
Category: sci.space
    Mean difference: 0.084, Same-category average similarity: 0.3
Category: misc.forsale
    Mean difference: 0.2, Same-category average similarity: 0.4
Category: soc.religion.christian
    Mean difference: 0.29, Same-category average similarity: 0.4
```

Challenges and Future extensions

- **Challenges**:
  - Inter category identification of similar documents
  - Manual loading of entire data set crashes
- **Future extensions**:
  - Use of better transformers for detecting granular level similarity
  - Extend this algorithm perfectly for plagiarism detection
  - Better processing of input data (eg: emotions, pos tag,punctuation)

Questions