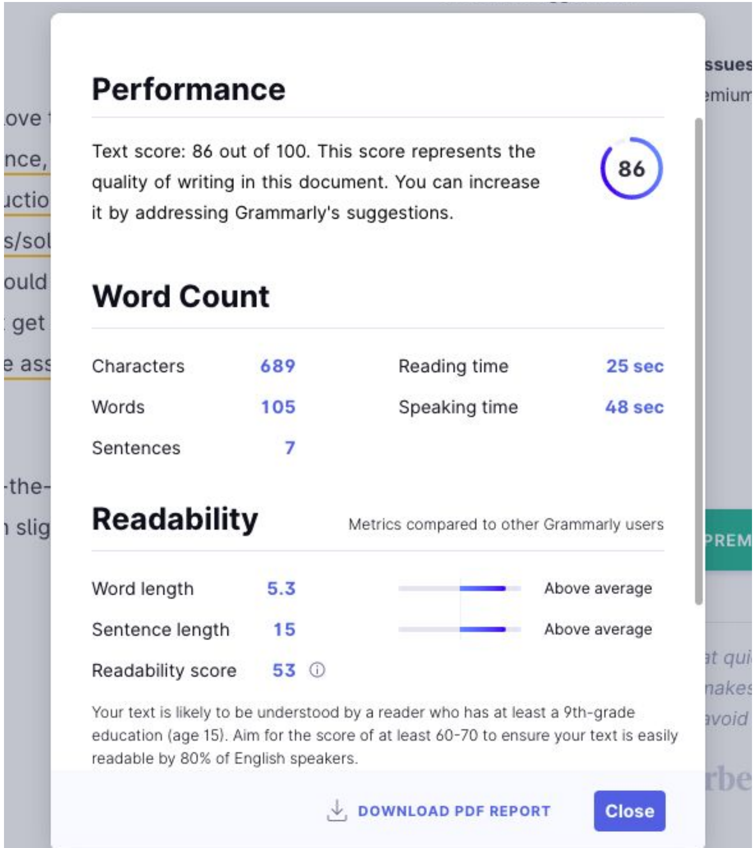
NLP + TDA

Parth Parikh

Spring 2022

Have you seen this on Grammarly?

Your text is likely to be understood by a reader who has at least a 9th-grade education (age 15). Aim for the score of at least 60-70 to ensure your text is easily readable by 80% of English speakers.



Performance

Text score: 86 out of 100. This score represents the quality of writing in this document. You can increase it by addressing Grammarly's suggestions.

86

Word Count

Characters	689	Reading time	25 sec
Words	105	Speaking time	48 sec
Sentences	7		

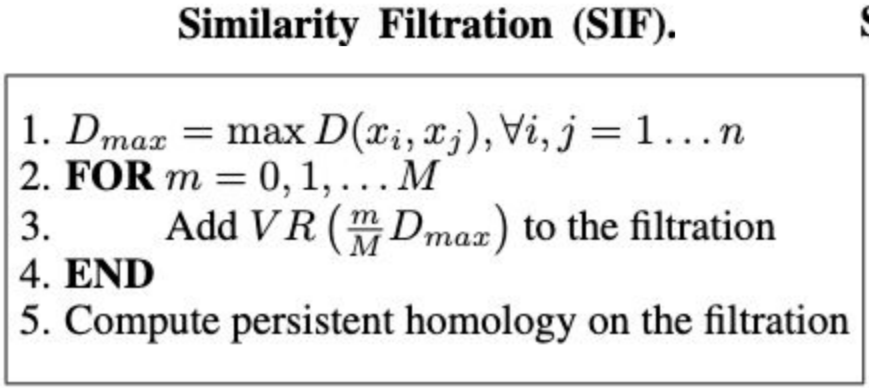
Readability

Metrics compared to other Grammarly users

Word length	5.3	Above average
Sentence length	15	Above average
Readability score	53 ⓘ	

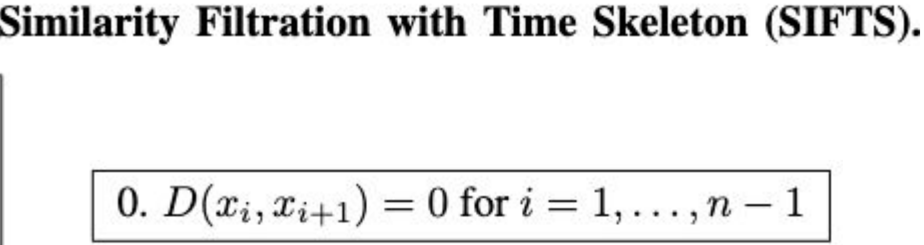
Your text is likely to be understood by a reader who has at least a 9th-grade education (age 15). Aim for the score of at least 60-70 to ensure your text is easily readable by 80% of English speakers.

Semantic *tie-backs* in a text document



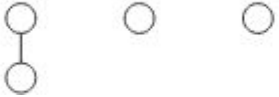
Similarity Filtration (SIF).

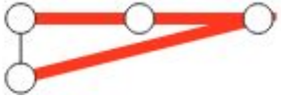
```
1. $D_{max} = \max D(x_i, x_j), \forall i, j = 1 \ldots n$
2. **FOR** $m = 0, 1, \ldots M$
3.         Add $VR\left(\frac{m}{M} D_{max}\right)$ to the filtration
4. **END**
5. Compute persistent homology on the filtration
```



Similarity Filtration with Time Skeleton (SIFTS).

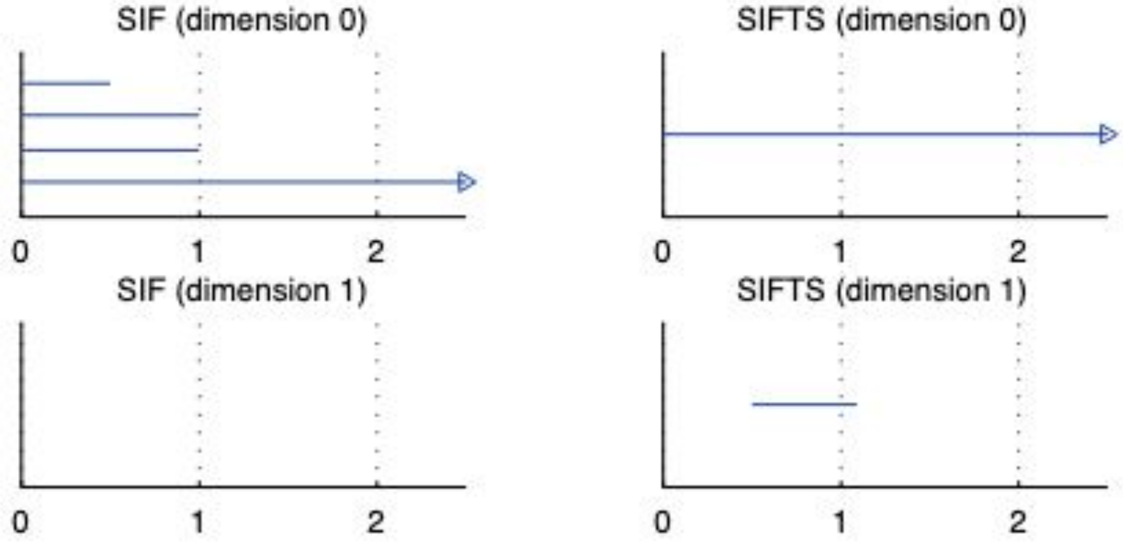
0. $D(x_i, x_{i+1}) = 0$ for $i = 1, \ldots, n-1$





Xiaojin Zhu UWM

Semantic *tie-backs* in a text document



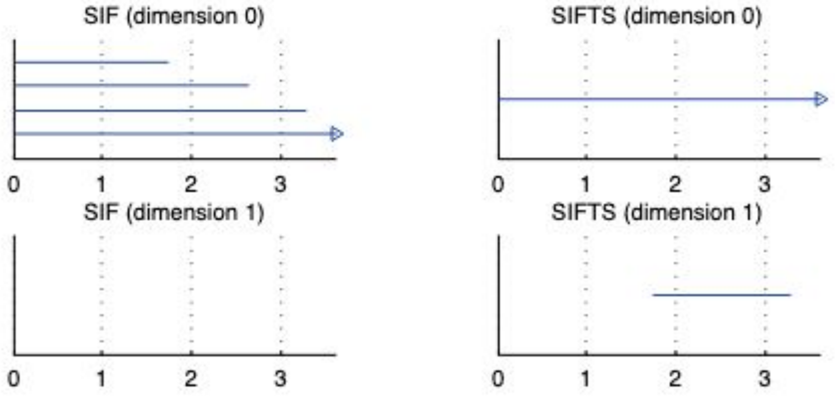
Xiaojin Zhu UWM

Semantic *tie-backs* in a text document - *Nursery Rhymes*

Euclidean distance between sentence-level bag-of-words count vectors

Filtrations has M = 100 steps

The itsy bitsy spider climbed up the waterspout.
Down came the rain
And washed the spider out.
Out came the sun
And dried up all the rain
And the itsy bitsy spider climbed up the spout again.

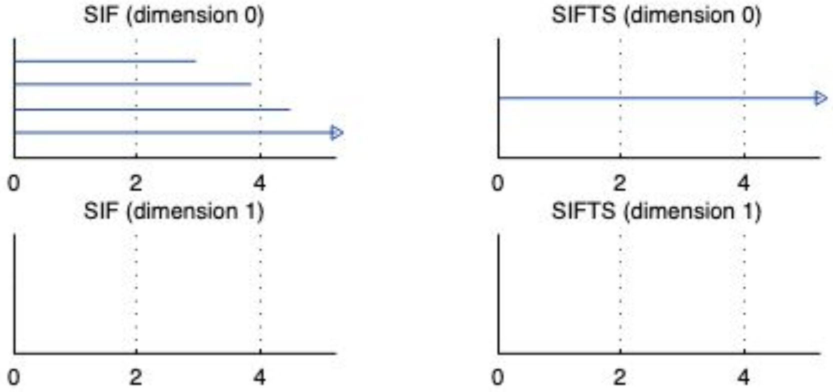


(a) Itsy Bitsy Spider

Xiaojin Zhu UWM

Semantic *tie-backs* in a text document - *Nursery Rhymes*

Row, row, row your boat
Gently down the stream
Merrily merrily, merrily, merrily
Life is but a dream



(b) Row Row Row Your Boat

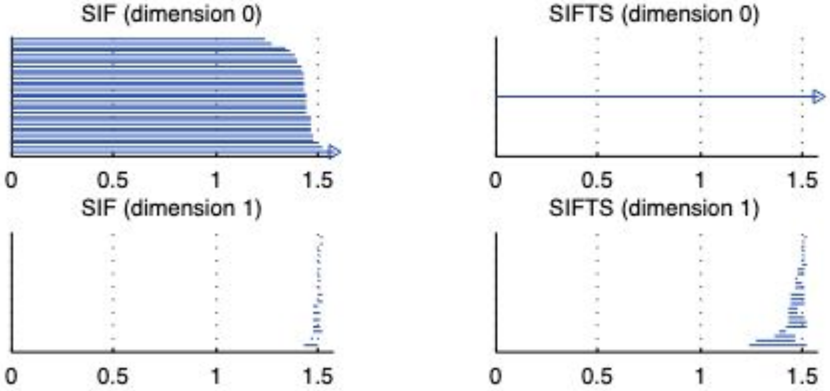
Xiaojin Zhu UWM

Semantic *tie-backs* in a text document - *Longer Documents*

Penn Treebank tokenization, case-folding, punctuation removal, and SMART stopword removal

Each text unit is converted to a tf.idf vector

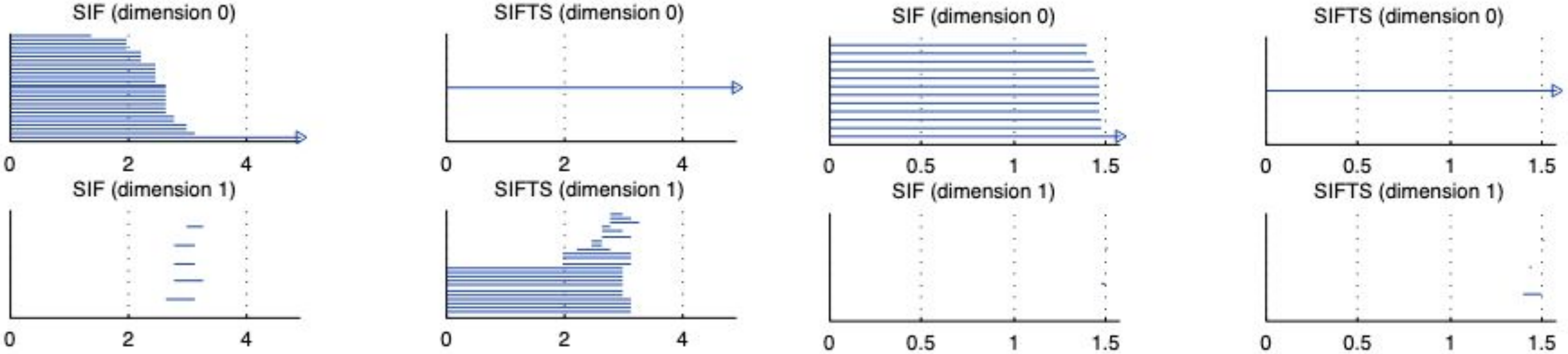
$$D(x_i, x_j) = \cos^{-1}\left(\frac{x_i^\top x_j}{\|x_i\| \cdot \|x_j\|}\right).$$



(d) The Emperor's New Clothes

Xiaojin Zhu UWM

Semantic *tie-backs* in a text document - *Longer Documents*



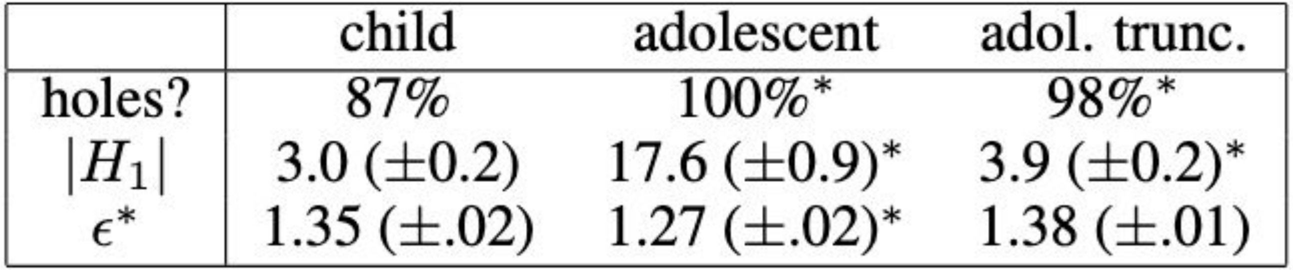
(c) London Bridge

(f) Alice in Wonderland

Xiaojin Zhu UWM

Semantic *tie-backs* in a text document - *Observations*

- *Older writers have more 1-homology groups than younger writers*
- $|H_1|$, the total number of 1st-order persistent homology classes (holes) over the whole range epsilon
  - counting the number of bars
- Epsilon-star - the smallest epsilon when the first hole in H_1 forms

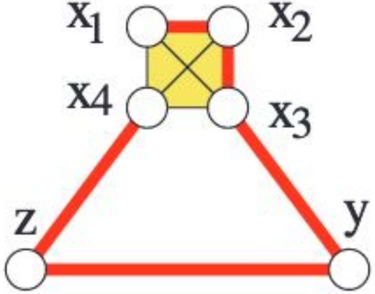


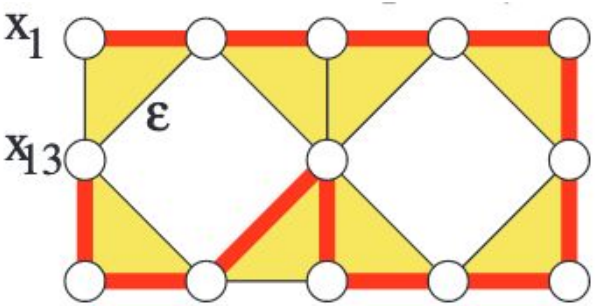
	child	adolescent	adol. trunc.
holes?	87%	100%*	98%*
$\|H_1\|$	3.0 (±0.2)	17.6 (±0.9)*	3.9 (±0.2)*
ϵ^*	1.35 (±.02)	1.27 (±.02)*	1.38 (±.01)

Xiaojin Zhu UWM

Semantic *tie-backs* in a text document - *Observations*

- *Homology is not just counting repeated text units*

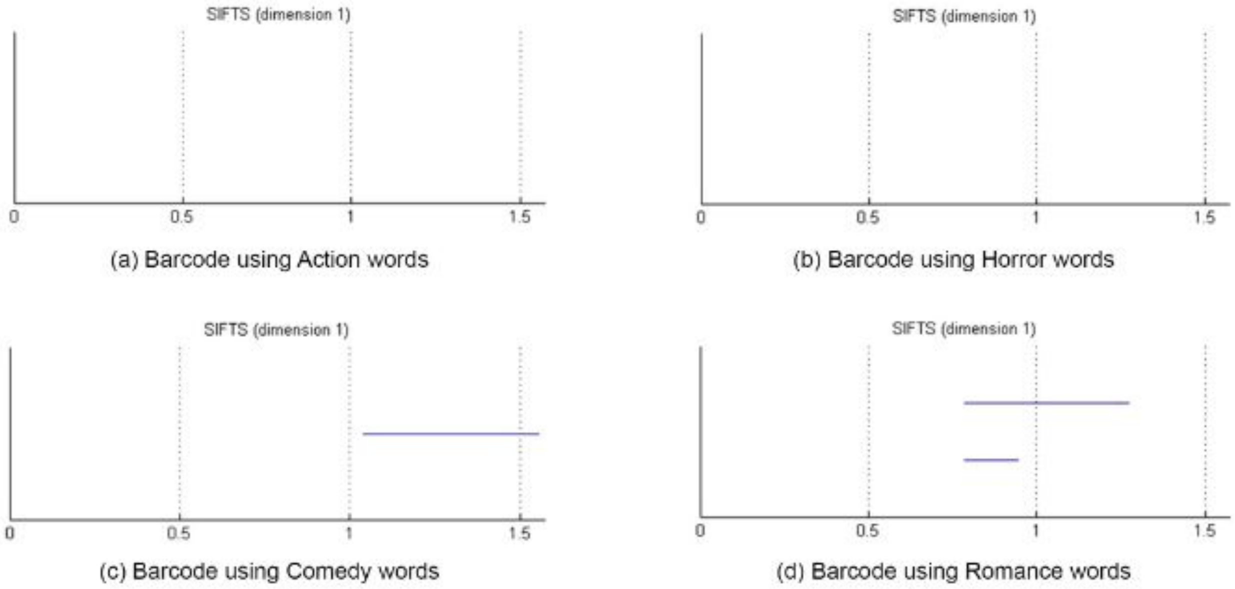




Xiaojin Zhu UWM

Movie Genre Detection Using Topological Data Analysis

- Predicting movie genres based on plot descriptions



(a) Barcode using Action words

(b) Barcode using Horror words

(c) Barcode using Comedy words

(d) Barcode using Romance words

Pratik Doshi UNC Charlotte

Movie Genre Detection Using Topological Data Analysis

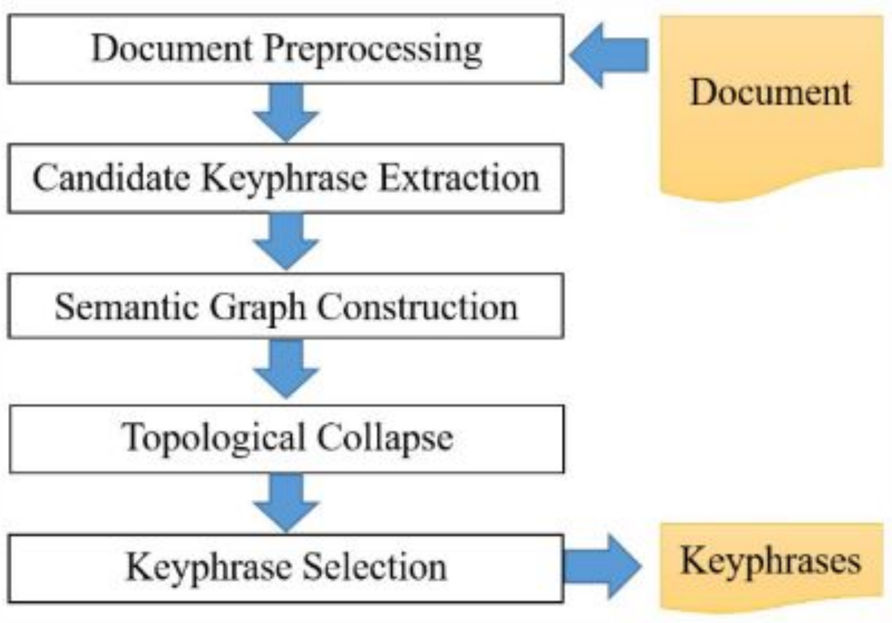
- Identify the top words using the TF-IDF measure
- Generate Term Frequency matrix for both the movie plots
- Find the 1-dimension holes across the sentences
    - Using the barcode representation of the 1-dimension homology complexes, the program is able to correctly identify the genres of 208 movies with overlapping genres, giving a hit rate of 0.8333%

Pratik Doshi UNC Charlotte

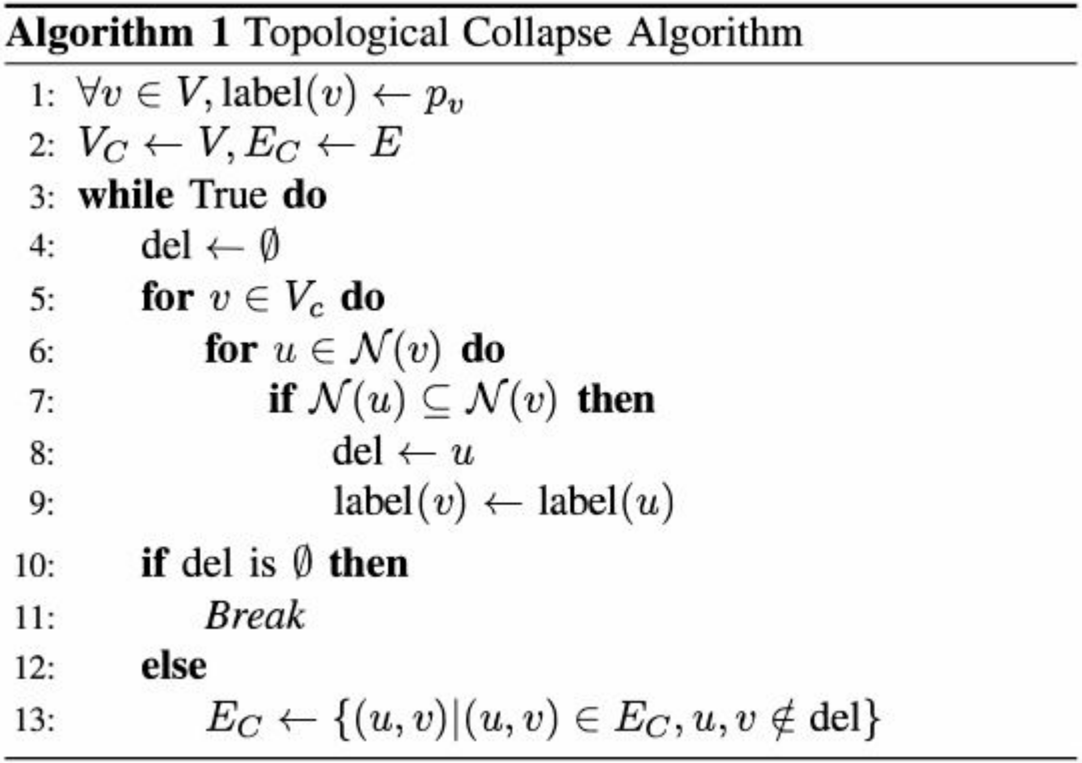
A Topological Collapse for Document Summarization

- DoCollapse: vertex dominance criterion
  - vertex v is dominated by vertex w if all vertices that share an edge with v also share an edge with vertex w
- Key Idea: *In a document semantic graph, if one candidate keyphrase dominates another one, then the dominating candidate should convey more important information and thus, is more likely to be a keyphrase*

A Topological Collapse for Document Summarization



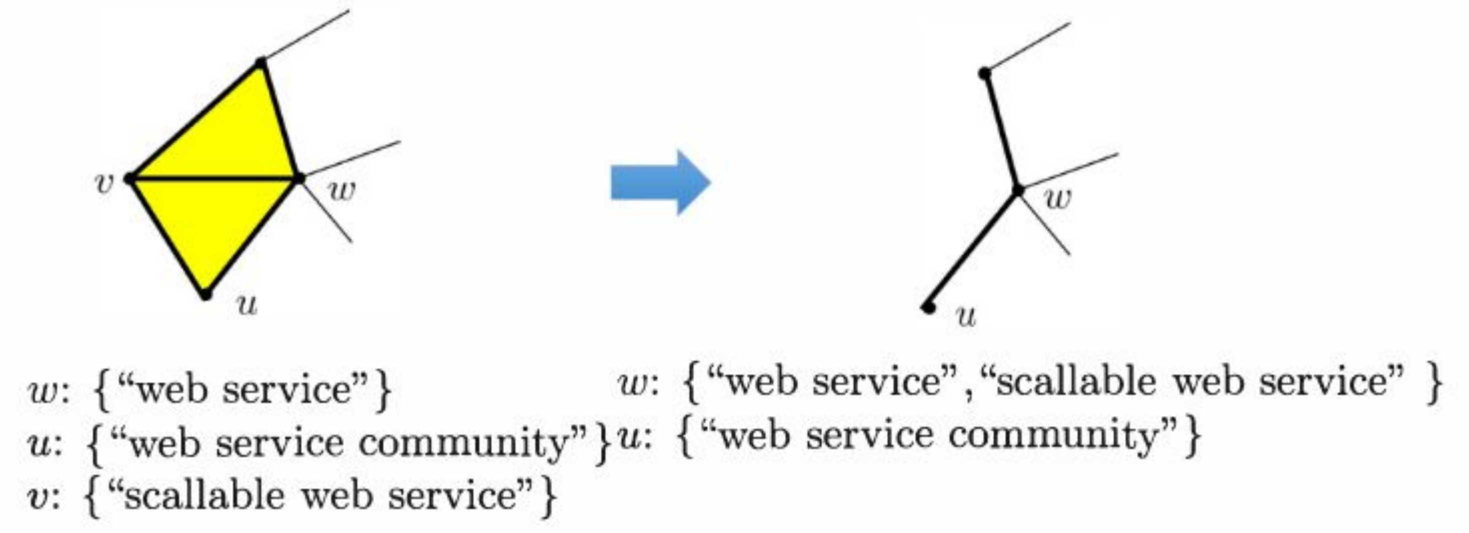
A Topological Collapse for Document Summarization



Algorithm 1 Topological Collapse Algorithm

```
$\forall v \in V, \text{label}(v) \leftarrow p_v$
$V_C \leftarrow V, E_C \leftarrow E$
while True do
    del $\leftarrow \emptyset$
    for $v \in V_c$ do
        for $u \in \mathcal{N}(v)$ do
            if $\mathcal{N}(u) \subseteq \mathcal{N}(v)$ then
                del $\leftarrow u$
                $\text{label}(v) \leftarrow \text{label}(u)$
    if del is $\emptyset$ then
        *Break*
    else
        $E_C \leftarrow \{(u, v) | (u, v) \in E_C, u, v \notin \text{del}\}$
```

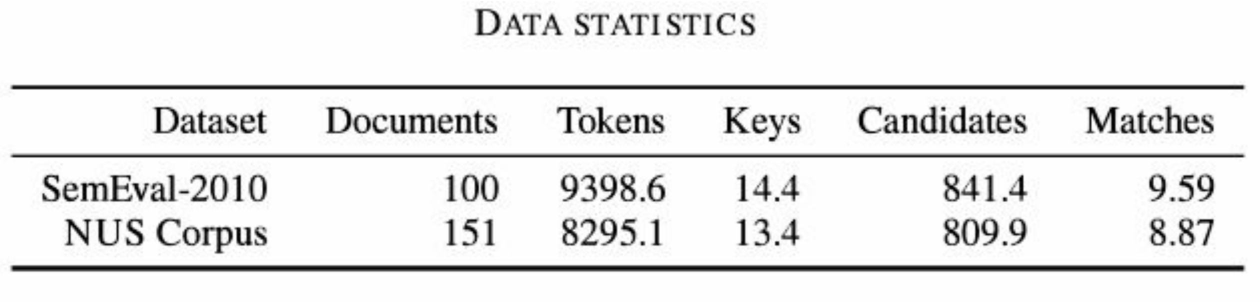
A Topological Collapse for Document Summarization



w: {“web service”}
u: {“web service community”}
v: {“scallable web service”}

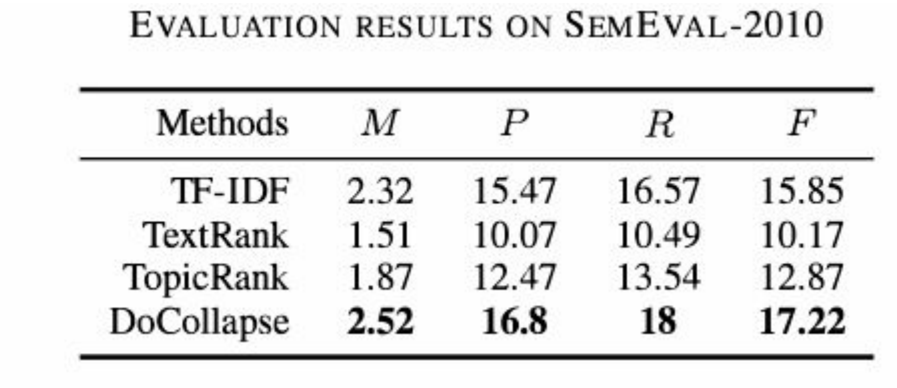
w: {“web service”,“scallable web service” }
u: {“web service community”}

A Topological Collapse for Document Summarization



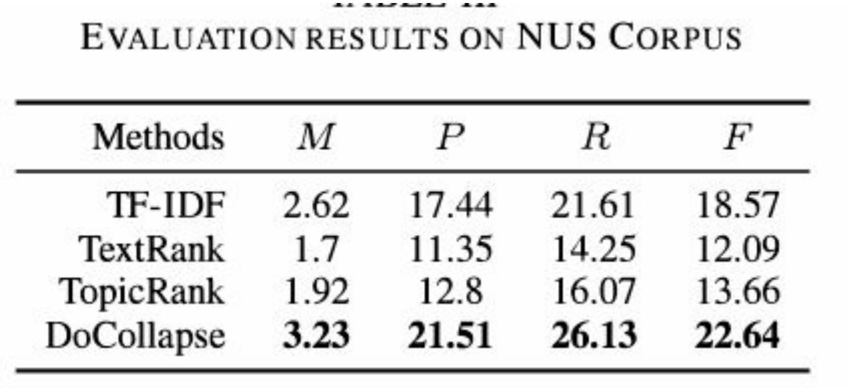
DATA STATISTICS

Dataset	Documents	Tokens	Keys	Candidates	Matches
SemEval-2010	100	9398.6	14.4	841.4	9.59
NUS Corpus	151	8295.1	13.4	809.9	8.87



EVALUATION RESULTS ON SEMEVAL-2010

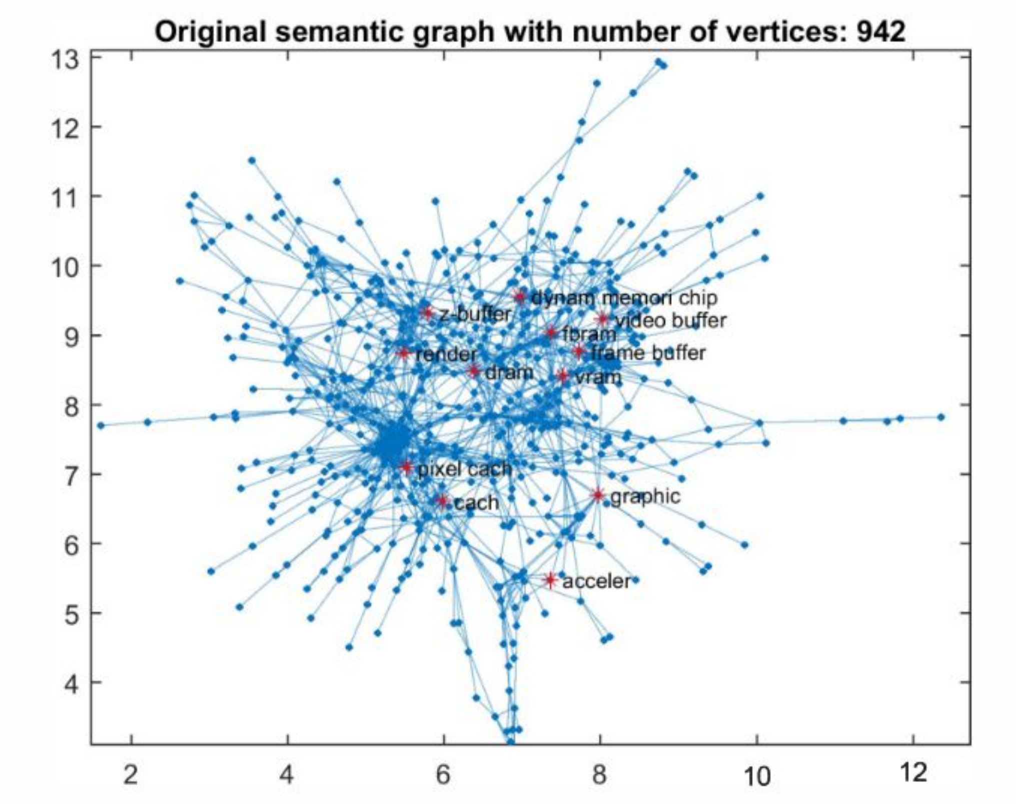
Methods	M	P	R	F
TF-IDF	2.32	15.47	16.57	15.85
TextRank	1.51	10.07	10.49	10.17
TopicRank	1.87	12.47	13.54	12.87
DoCollapse	**2.52**	**16.8**	**18**	**17.22**

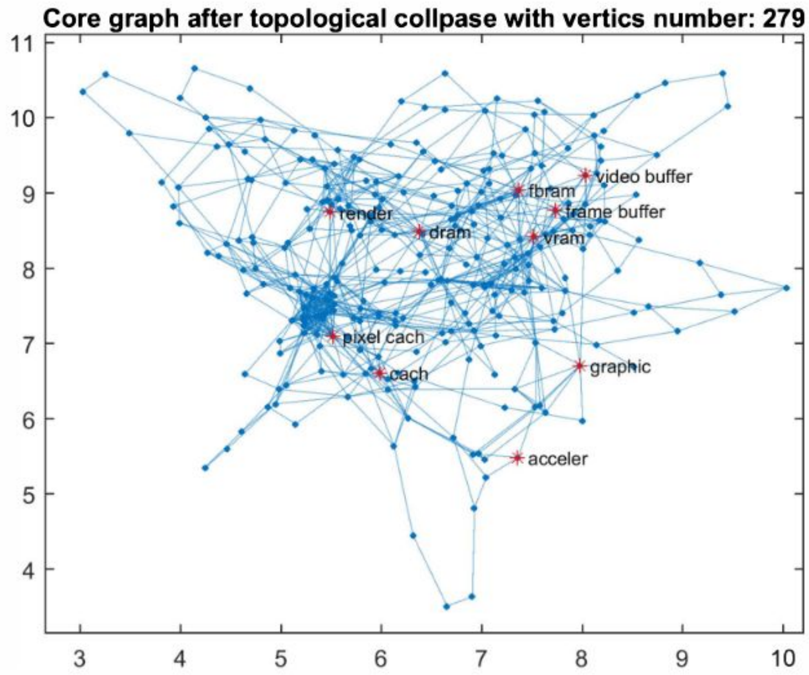


EVALUATION RESULTS ON NUS CORPUS

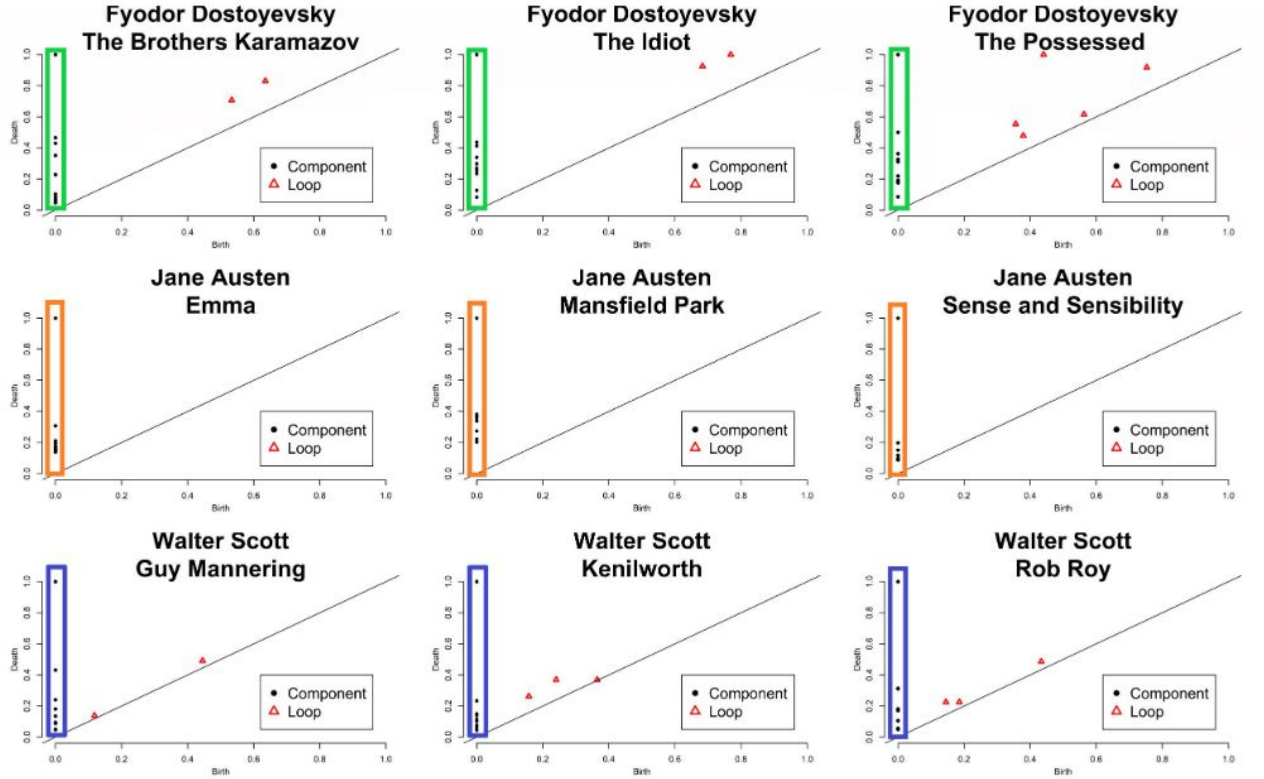
Methods	M	P	R	F
TF-IDF	2.62	17.44	21.61	18.57
TextRank	1.7	11.35	14.25	12.09
TopicRank	1.92	12.8	16.07	13.66
DoCollapse	**3.23**	**21.51**	**26.13**	**22.64**

A Topological Collapse for Document Summarization





Topological Signature of 19th Century Novelists - a skim



Fyodor Dostoyevsky The Brothers Karamazov

Fyodor Dostoyevsky The Idiot

Fyodor Dostoyevsky The Possessed

Jane Austen Emma

Jane Austen Mansfield Park

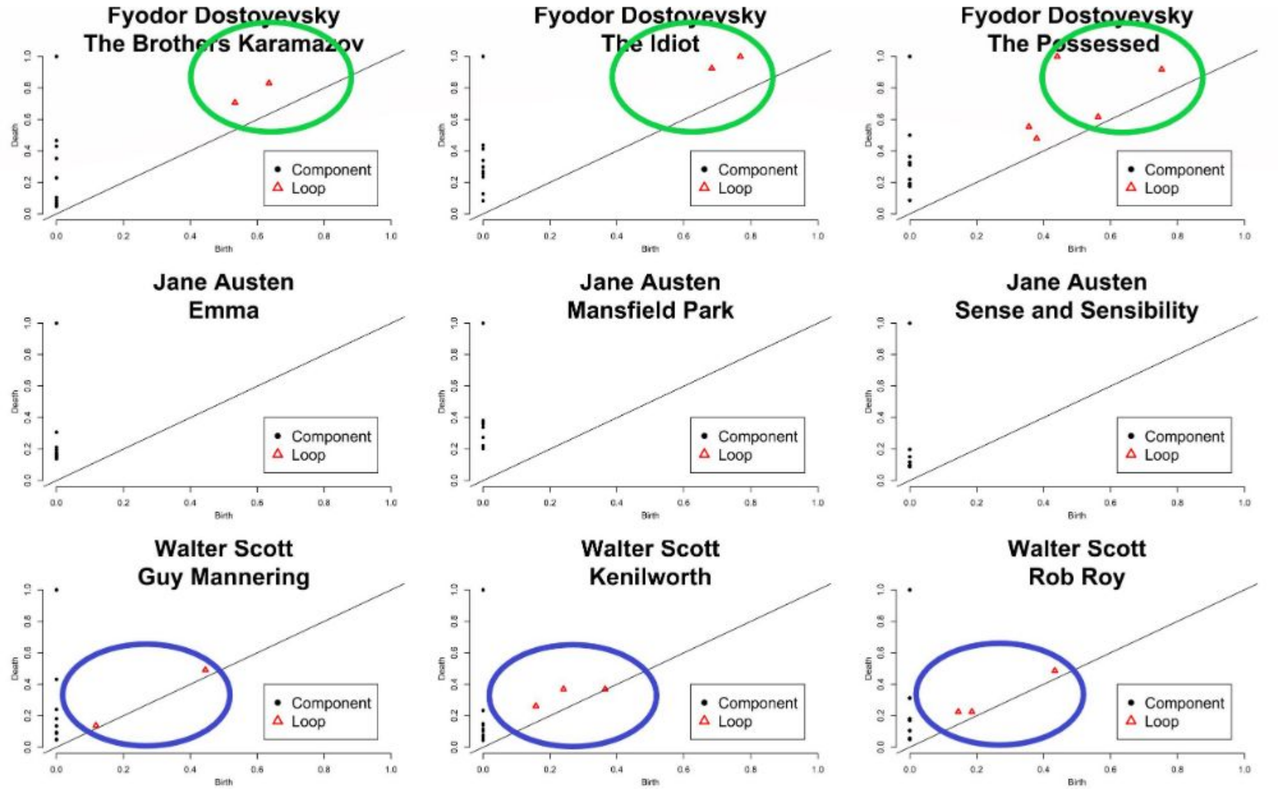
Jane Austen Sense and Sensibility

Walter Scott Guy Mannering

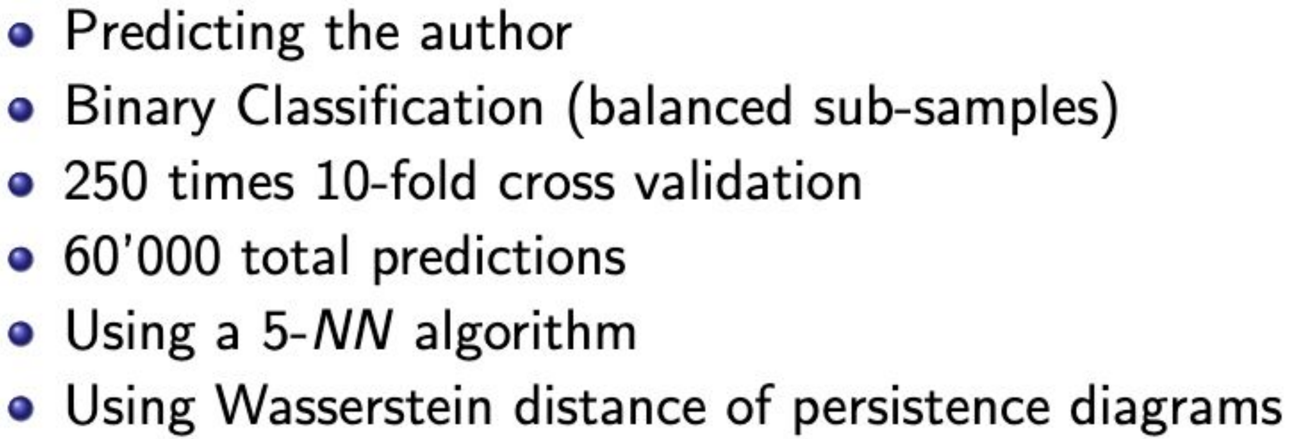
Walter Scott Kenilworth

Walter Scott Rob Roy

Topological Signature of 19th Century Novelists - a skim



Topological Signature of 19th Century Novelists - a skim



- Predicting the author
- Binary Classification (balanced sub-samples)
- 250 times 10-fold cross validation
- 60’000 total predictions
- Using a 5-*NN* algorithm
- Using Wasserstein distance of persistence diagrams

End