A Pipeline for Scalable Text Reuse Analysis

Milad Alshomary

Bauhaus Universität

05.07.2018
Overview

● **Motivation**

● **A Pipeline for Scalable Text Reuse Extraction**

● **Application on Wikipedia**

● **Application on Wikipedia and Common Crawl**

● **Conclusion**
Text Reuse (TR)

- Quoting
- Verbatim
- Paraphrasing
- Translation
- Summarization
TR Detection Applications

**Motivation**

METER project (Measuring Text Reuse)

Plagiarism detection

Although tooth eruption occurs at different times for different people, a general eruption timeline exists. Typically, humans have 20 primary teeth and 32 permanent teeth. The dentition goes through three stages: the first, mixed primary dentition, the second, mixed primary dentition, and the third, permanent dentition. The first permanent tooth erupts into the mouth, the tooth is visible, and the primary tooth falls out. After the last primary tooth is lost, the second permanent tooth erupts.
TR Detection Applications

Motivation

Plagiarism detection

METER projet (Measuring Text Reuse)
TR Detection Applications

Motivation

METER projet (Measuring Text Reuse)

Plagiarism detection

Although tooth eruption occurs at different times for different people, a general eruption timeline exists. Typically, humans have 28 primary teeth and 32 permanent teeth. The dentition grows through three stages: the first two stages of primary tooth eruption occur when the primary teeth are visible, and the third stage when the first permanent teeth erupt to the mouth. The teeth continue to develop in the mixed dentition stage. After the last primary tooth is erupted, the permanent teeth erupt to the mouth.
Wikipedia vs The World

- Digital Encyclopedia
- Collaborative environment
- Giant public source of information
- Free to use
Wikipedia vs The World

- Digital Encyclopedia
- Collaborative environment
- Giant public source of information
- Free to use
Wikipedia vs The World

- Digital Encyclopedia
- Collaborative environment
- Giant public source of information
- Free to use
Wikipedia vs The World

- Digital Encyclopedia
- Collaborative environment
- Giant public source of information
- Free to use
Motivation

Wikipedia vs The World

Quality Flaws → Scientific community

reddit → Scientific community

stack overflow → Scientific community
Wikipedia vs The World

- Web pages = Wikipedia text + advertisements
Research Questions

➔ What kinds of text reuse occur within Wikipedia?
➔ How much of the web is a copy of Wikipedia content?
➔ How much revenue does this content generate?
Research Questions

➔ What kinds of text reuse occur within Wikipedia?
➔ How much of the web is a copy of Wikipedia content?
➔ How much revenue does this content generate?
Research Questions

➔ What kinds of text reuse occur within Wikipedia?
➔ How much of the web is a copy of Wikipedia content?
➔ How much revenue does this content generate?
A Pipeline for Scalable Text Reuse Extraction
A Pipeline for Scalable Text Reuse Extraction

Text Reuse Pipeline

D1 ➔ TR Pipeline ➔ D2

➔ Input: Two datasets
➔ Output: Text reuse cases
Text Reuse Pipeline

→ Input: Two datasets
→ Output: Text reuse cases
Text Reuse Pipeline

→ Content extraction
→ Chunking
→ Feature extraction

A Pipeline for Scalable Text Reuse Extraction
Text Reuse Pipeline

A Pipeline for Scalable Text Reuse Extraction

Text Preprocessing ➔ Content extraction ➔ Chunking ➔ Feature extraction

Candidate Elimination ➔ Pairwise scan ➔ Text Reuse heuristics

Text Alignment
Text Reuse Pipeline

- Text Preprocessing → Content extraction
- Text Preprocessing → Chunking
- Text Preprocessing → Feature extraction
- Candidate Elimination → Pairwise scan
- Candidate Elimination → Text Reuse heuristics
- Text Alignment → Detailed scan of text reuse
- Text Alignment → Picapica framework
Candidate Elimination

Keys for scaling-up:

➔ Cluster computing
➔ Heuristics based candidate elimination algorithms
Candidate Elimination

Keys for scaling-up:

➔ Cluster computing
➔ Heuristics based candidate elimination algorithms
Candidate Elimination

For a *candidacy* function we proposed the following methods:

- Cosine similarity of TF-IDF (semantic)
- Paragraph embedding (semantic)
- Stopwords N-grams (structure)
- Weighted average of Stopwords Ngrams and Paragraph embedding (semantic + structure)
Candidate Elimination

Generate TR Sample from Wikipedia:
- Sample 1k documents from Wikipedia
- Using Picapica framework to find TR cases
Candidate Elimination

Generate TR Sample from Wikipedia:
- Sample 1k documents from Wikipedia
- Using Picapica framework to find TR cases
A Pipeline for Scalable Text Reuse Extraction

Candidate Elimination

Generate TR Sample from Wikipedia:
- Sample 1k documents from Wikipedia
- Using Picapica framework to find TR cases

- 232 documents
- ~ 90% have < 10 alignments (TR case)
Candidate Elimination

Evaluation of “candidacy” function:

- For each document in TR sample:
  - Sort all Wikipedia articles according to the proposed “candidacy”.
  - Precision/Recall on Thresholds of [1, 101,...,100k]
  - A True Positive (TP) is a pair of documents that have TR.
Candidate Elimination

Evaluation of “candidacy” function:

![Graph showing evaluation of candidacy function with precision and recall metrics.]

- \( t1 \)
- \( t2 \)
- \( t3 \)
- \( r1 \)
- \( r2 \)
- \( p1 \)
- \( p2 \)

A Pipeline for Scalable Text Reuse Extraction
Candidate Elimination

Semantic hashing function:

- Hashes documents into binary hashes.
- Similar documents get similar or exact binary hash.
Candidate Elimination

Semantic hashing function:
- Hashing all documents.
- Inverted index.
- Hash document’s chunks.
- Apply *candidacy* function only on documents that intersect in one hash at least.
Candidate Elimination

Semantic hashing function:

- Hashing all documents.
- Inverted index.
- Hash document’s chunks.
- Apply candidacy function only on documents that intersect in one hash at least.
Candidate Elimination

Semantic hashing function:
- Hashing all documents.
- Inverted index.
- Hash document’s chunks.
- Apply candidacy function only on documents that intersect in one hash at least.
Candidate Elimination

Semantic hashing function:
- Hashing all documents.
- Inverted index.
- Hash document’s chunks.
- Apply *candidacy* function only on documents that intersect in one hash at least.
Candidate Elimination

Proposed semantic hashing methods:

- Random Projection (data independent)
- Variational Deep Semantic Hashing (data dependent)
Candidate Elimination

Proposed semantic hashing methods:

- Random Projection (data independent)
- Variational Deep Semantic Hashing (data dependent)
Candidate Elimination

Proposed semantic hashing methods:

- Random Projection (data independent)
- Variational Deep Semantic Hashing (data dependent)
Candidate Elimination

Proposed semantic hashing methods:
- Random Projection (data independent)
- Variational Deep Semantic Hashing (data dependent)
Candidate Elimination

Hashing methods evaluation:

- Using same TR sample for evaluation.
- Hashing all documents using the proposed hashing function.
- Compute precision and recall.
Candidate Elimination

Hashing methods evaluation:

- Using same TR sample for evaluation.
- Hashing all documents using the proposed hashing function.
- Compute precision and recall.
Candidate Elimination

Hashing methods evaluation:
- Using same TR sample for evaluation.
- Hashing all documents using the proposed hashing function.
- Compute precision and recall.

Precision = 2/3
Recall = 1.0
Candidate Elimination

Hashing methods evaluation

- Using same TR sample for evaluation.
- Hashing all documents using the proposed hashing function.
- Compute precision and recall.

<table>
<thead>
<tr>
<th>Random projection</th>
<th>bits</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>....</td>
<td>8</td>
<td>3.1 x 10^{-4}</td>
</tr>
<tr>
<td></td>
<td>....</td>
<td>16</td>
<td>9.9 x 10^{-4}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VDSH</th>
<th>bits</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>....</td>
<td>8</td>
<td>2.8 x 10^{-4}</td>
<td>0.88</td>
</tr>
<tr>
<td>....</td>
<td>16</td>
<td>4.5 x 10^{-3}</td>
<td>0.73</td>
</tr>
</tbody>
</table>
Candidate Elimination

Hashing methods evaluation

- Using same TR sample for evaluation.
- Hashing all documents using the proposed hashing function.
- Compute precision and recall.

<table>
<thead>
<tr>
<th>VDSH</th>
<th>bits</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8</td>
<td>$2.8 \times 10^{-4}$</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>$4.5 \times 10^{-3}$</td>
<td>0.73</td>
</tr>
</tbody>
</table>

- Retains 73% of the recall
- By experiment:
  - Reduces the computations needed by $3$ order of magnitude
Application on Wikipedia
Text Reuse In Wikipedia

➔ What kinds of text reuse occur within Wikipedia?
➔ How much of the web is a copy of Wikipedia content?
➔ How much revenue does this content generate?
Text Reuse In Wikipedia

Wikipedia

TR Pipeline

100 million text reuse

360k Wikipedia Article

Contains TR (9%) Wikipedia Articles

Pipeline for TR extraction

Milad Alshomary

05.07.2018
What kinds of text reuse occur in Wikipedia?

- Reasons behind text reuse:
  (1) Two texts describe the same topic.
  (2) Two texts describe two different topics, that share similar characteristics.
What kinds of text reuse occur in Wikipedia?

- Reasons behind text reuse:
  (1) Two texts describe the same topic.
Text Reuse In Wikipedia

What kinds of text reuse occur in Wikipedia?

- Reuse in Structure

Tooth eruption

Although tooth eruption occurs at different times for different people, a general eruption timeline exists. Typically, humans have 20 primary teeth and 32 permanent teeth. The dentition goes through three stages. The first, known as primary dentition stage, occurs when only primary teeth are visible. Once the first permanent tooth erupts into the mouth, the teeth that are visible are in the mixed (or transitional) dentition stage. After the last primary tooth is shed or exfoliates out of the mouth,

Human tooth development

Although tooth eruption occurs at different times for different people, a general eruption timeline exists. Typically, humans have 20 primary (baby) teeth and 32 permanent teeth. Tooth eruption has three stages. The first, known as deciduous dentition stage, occurs when only primary teeth are visible. Once the first permanent tooth erupts into the mouth, the teeth are in the mixed (or transitional) dentition. After the last primary tooth falls out of the mouth, a process known as
Text Reuse In Wikipedia

What kinds of text reuse occur in Wikipedia?

- Reasons behind text reuse:
  
  (2) Two texts describe two different topics, that share similar characteristics

[Diagram showing text reuse with examples from Wikipedia pages for Zimna Woda and Niedźwiedzie.]
What kinds of text reuse occur in Wikipedia?

- Reasons behind text reuse:
  (2) Two texts describe two different topics, that share similar characteristics.
Text Reuse In Wikipedia

- Vertical alignment $\rightarrow$ Content TR
- Horizontal alignment $\rightarrow$ Structure TR
Text Reuse In Wikipedia

- Vertical alignment → Content TR
- Horizontal alignment → Structure TR
Text Reuse In Wikipedia

- Vertical alignment → Content TR
- Horizontal alignment → Structure TR
Application on Wikipedia and Common Crawl
Wikipedia vs The Web

➔ What kinds of text reuse occur within Wikipedia?
➔ How much of the web is a copy of Wikipedia content?
➔ How much revenue does this content generate?
Wikipedia vs The Web

WWW → Common Crawl
- Crawling

Common Crawl → Extracted web content
- Content extraction
- Keeping only english pages

Extracted web content → 10% random sample
Web Sample

Application on Wikipedia and Common Crawl
Wikipedia vs The Web

- **Crawling**

- **Content extraction**
  - Keeping only English pages

- **Extracted web content**

- **Web Sample**
  - 10% random sample

- **Number of websites**
  - 59 million web pages.
  - 1.4 million websites.
  - 70% of these websites contain less than 10 web pages
Wikipedia vs The Web

- 1.6 million text reuse cases.
- 15k pages reuse Wikipedia text.
- 4.8k websites reuse Wikipedia text.

Page with TR from Wikipedia (2.5%)
Wikipedia vs The Web

Monthly revenue estimation:

- Rough estimate of Ads revenue
- Based on CPM (Cost Per Millie)
- Sampled 100 webpages and manually checked the existence of Advertisements.
Revenue estimation:

- Per website (all websites)
- Per website (highly reusing)
- Per Wikipedia web page

<table>
<thead>
<tr>
<th>website</th>
<th>Monthly revenue</th>
<th>Percentage of reuse</th>
<th>Monthly Wikipedia value</th>
</tr>
</thead>
<tbody>
<tr>
<td>pdxretro.com</td>
<td>$195</td>
<td>0.012</td>
<td>$2.5</td>
</tr>
<tr>
<td>seqrchquarry.com</td>
<td>$8,850</td>
<td>0.096</td>
<td>$850</td>
</tr>
<tr>
<td>asiatees.com</td>
<td>$36,000</td>
<td>0.017</td>
<td>$613</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>$1.2 million</td>
</tr>
</tbody>
</table>
Wikipedia vs The Web

Revenue estimation:

- Per website (all websites)
- Per website (highly reusing)
- Per Wikipedia web page

<table>
<thead>
<tr>
<th>website</th>
<th>Monthly revenue</th>
<th>Percentage of reuse</th>
<th>Monthly Wikipedia value</th>
</tr>
</thead>
<tbody>
<tr>
<td>pdxretro.com</td>
<td>$195</td>
<td>0.012</td>
<td>$2.5</td>
</tr>
<tr>
<td>seqrchquarry.com</td>
<td>$8,850</td>
<td>0.096</td>
<td>$850</td>
</tr>
<tr>
<td>asiatees.com</td>
<td>$36,000</td>
<td>0.017</td>
<td>$613</td>
</tr>
<tr>
<td>.....</td>
<td>.....</td>
<td>.....</td>
<td>.....</td>
</tr>
<tr>
<td>Total</td>
<td>.....</td>
<td>.....</td>
<td>$1.2 million</td>
</tr>
</tbody>
</table>
Wikipedia vs The Web

Revenue estimation:

- Per website (all websites)
- Per website (highly reusing)
- Per Wikipedia web page

<table>
<thead>
<tr>
<th>Website</th>
<th>Monthly revenue</th>
<th>Percentage of reuse</th>
<th>Monthly Wikipedia value</th>
</tr>
</thead>
<tbody>
<tr>
<td>pdxretro.com</td>
<td>$195</td>
<td>0.012</td>
<td>$2.5</td>
</tr>
<tr>
<td>seqrchquarry.com</td>
<td>$8,850</td>
<td>0.096</td>
<td>$850</td>
</tr>
<tr>
<td>asiatees.com</td>
<td>$36,000</td>
<td>0.017</td>
<td>$613</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>$1.2 million</td>
</tr>
</tbody>
</table>
Wikipedia vs The Web

Revenue estimation:
- Per website (all websites)
- Per website (highly reusing)
- Per Wikipedia web page

<table>
<thead>
<tr>
<th>website</th>
<th>Monthly revenue</th>
<th>Percentage of reuse</th>
<th>Monthly Wikipedia value</th>
</tr>
</thead>
<tbody>
<tr>
<td>pdxretro.com</td>
<td>$195</td>
<td>0.012</td>
<td>$2.5</td>
</tr>
<tr>
<td>seqrchquary.com</td>
<td>$8,850</td>
<td>0.096</td>
<td>$850</td>
</tr>
<tr>
<td>asiatees.com</td>
<td>$36,000</td>
<td>0.017</td>
<td>$613</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>$1.2 million</td>
</tr>
</tbody>
</table>
Wikipedia vs The Web

Revenue estimation:

- Per website (all websites)
- Per website (highly reusing)
- Per Wikipedia web page

<table>
<thead>
<tr>
<th>website</th>
<th>Monthly revenue</th>
<th>Percentage of reuse</th>
<th>Monthly Wikipedia value</th>
</tr>
</thead>
<tbody>
<tr>
<td>pdxretro.com</td>
<td>$195</td>
<td>0.012</td>
<td>$2.5</td>
</tr>
<tr>
<td>seqrchquarry.com</td>
<td>$8,850</td>
<td>0.096</td>
<td>$850</td>
</tr>
<tr>
<td>asiatees.com</td>
<td>$36,000</td>
<td>0.017</td>
<td>$613</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td><strong>$1.2 million</strong></td>
</tr>
</tbody>
</table>

The rough estimate of monthly revenue of Wikipedia content
Wikipedia vs The Web

Revenue estimation:

- Per website (all websites)
- Per website (highly reusing)
- Per Wikipedia web page

- Percentage of pages reusing Wikipedia >= 0.5
- 87 websites.
- Estimated monthly revenue: $15k
Wikipedia vs The Web

Revenue estimation:
- Per website (all websites)
- Per website (highly reusing)
- Per Wikipedia web page

<table>
<thead>
<tr>
<th>Reused Wikipedia page</th>
<th>Average page views</th>
<th>Average CPM</th>
<th>Average monthly revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuclear renaissance</td>
<td>645</td>
<td>$2.8</td>
<td>$1.806</td>
</tr>
<tr>
<td>Second Chechen War</td>
<td>34655</td>
<td>$2.8</td>
<td>$97</td>
</tr>
<tr>
<td>Enumerated powers</td>
<td>12858</td>
<td>$2.8</td>
<td>$36</td>
</tr>
<tr>
<td>...</td>
<td>.....</td>
<td>.....</td>
<td>.....</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>$900k</td>
</tr>
</tbody>
</table>
## Revenue estimation:

- Per website (all websites)
- Per website (highly reusing)
- Per Wikipedia web page

<table>
<thead>
<tr>
<th>Reused Wikipedia page</th>
<th>Average page views</th>
<th>Average CPM</th>
<th>Average monthly revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuclear renaissance</td>
<td>645</td>
<td>$2.8</td>
<td>$1.806</td>
</tr>
<tr>
<td>Second Chechen War</td>
<td>34655</td>
<td>$2.8</td>
<td>$97</td>
</tr>
<tr>
<td>Enumerated powers</td>
<td>12858</td>
<td>$2.8</td>
<td>$36</td>
</tr>
<tr>
<td></td>
<td>.....</td>
<td>.....</td>
<td>.....</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>$900k</td>
</tr>
</tbody>
</table>

Estimated from marketing reports
Wikipedia vs The Web

Revenue estimation:

- Per website (all websites)
- Per website (highly reusing)
- Per Wikipedia web page

<table>
<thead>
<tr>
<th>Reused Wikipedia page</th>
<th>Average page views</th>
<th>Average CPM</th>
<th>Average monthly revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuclear renaissance</td>
<td>645</td>
<td>$2.8</td>
<td>$1.806</td>
</tr>
<tr>
<td>Second Chechen War</td>
<td>34655</td>
<td>$2.8</td>
<td>$97</td>
</tr>
<tr>
<td>Enumerated powers</td>
<td>12858</td>
<td>$2.8</td>
<td>$36</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>$900k</td>
</tr>
</tbody>
</table>
Wikipedia vs The Web

Revenue estimation:

- Per website (all websites)
- Per website (highly reusing)
- Per Wikipedia web page

<table>
<thead>
<tr>
<th>Reused Wikipedia page</th>
<th>Average page views</th>
<th>Average CPM</th>
<th>Average monthly revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuclear renaissance</td>
<td>645</td>
<td>$2.8</td>
<td>$1.806</td>
</tr>
<tr>
<td>Second Chechen War</td>
<td>34655</td>
<td>$2.8</td>
<td>$97</td>
</tr>
<tr>
<td>Enumerated powers</td>
<td>12858</td>
<td>$2.8</td>
<td>$36</td>
</tr>
<tr>
<td>...</td>
<td>.....</td>
<td>.....</td>
<td>.....</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>$900k</td>
</tr>
</tbody>
</table>
### Monthly revenue:

<table>
<thead>
<tr>
<th>Per Web sample</th>
<th>Number of reusing web pages</th>
<th>Revenue(per webpage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>59 million</td>
<td>15k</td>
<td>$900k</td>
</tr>
<tr>
<td>590 million</td>
<td>150k</td>
<td>$9 million</td>
</tr>
</tbody>
</table>
## Wikipedia vs The Web

### Monthly revenue:

<table>
<thead>
<tr>
<th>Per Web sample</th>
<th>Number of reusing web pages</th>
<th>Revenue (per webpage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>59 million</td>
<td>15k</td>
<td>$900k</td>
</tr>
<tr>
<td>590 million</td>
<td>150k</td>
<td>$9 million</td>
</tr>
</tbody>
</table>

Application on Wikipedia and Common Crawl
Conclusion
Summary

- Pipeline for TR Extraction
- Text Reuse in Wikipedia
- Text Reuse between Wikipedia and the Web

<table>
<thead>
<tr>
<th>Per website (all websites)</th>
<th>Per website (highly reuse)</th>
<th>Per Webpage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1.2 million</td>
<td>$15k</td>
<td>$900k</td>
</tr>
</tbody>
</table>
Future Work

- Using the pipeline to extract and analyze TR between Wikipedia and the scientific community.
- Experiments on the Text Alignment subtask.
- Further analysis of the extracted Text Reuse cases.
Conclusion

- Using the pipeline to extract and analyze TR between Wikipedia and the scientific community.
- Experiments on the Text Alignment subtask.
- Further analysis of the extracted Text Reuse cases.
Conclusion

- Using the pipeline to extract and analyze TR between Wikipedia and the scientific community.
- Experiments on the Text Alignment subtask.
- Further analysis of the extracted Text Reuse cases.
Conclusion

- Using the pipeline to extract and analyze TR between Wikipedia and the scientific community.
- Experiments on the Text Alignment subtask.
- Further analysis of the extracted Text Reuse cases.
Backup Slides
- **Candidate Elimination functions:**

\[
\text{cand}(s, d_i) = \max(\text{cosine\_similarity}(p_i, p_j)) : p_i \in d_i, p_j \in d_i
\]

\[
\text{cand}(s, d_i) = \max(\frac{\text{count}(\text{shared\_ngrams}(p_i, p_j))}{\min(\text{ngrams\_count}(p_i), \text{ngrams\_count}(p_j))}) : p_i \in d_i, p_j \in d_i
\]
- **Stopwords N-grams procedure:**

  **Wiki paragraphs** → **stopwords** → **stopword ngrams** → **filtered stopword ngrams** → **binary count vector**

  **Top 50 frequent stopwords:**
  the, of, and, a, in, to, is, was, it, for, with, he, be, on, i, that, by, at, you, 's, are, not, his, this, from, but, had, which, she, they, or, on, were, we, their, been, has, have, will, would, her, there, can, all, as, if, who, what, said

  - Let $C = \{\text{the, of, and, a, in, to, 's}\}$ stopwords that increases false positive.

  - $X$ is accepted n-gram if:
    - It doesn't contain more than n-1 stopwords from C
    - The maximal sequence of stopwords belonging to C is less than n-2

  - Binary count vector ignores the frequency in which a specific n-gram happened in a paragraph.

  - We apply the scoring function on the binary count vector
- VDSH explained:

VDSH USAGE

Train

Wikipedia

Model fitting

Top 260k frequent words

Web

IDF

Transform

Wikipedia

tf-idf model

wiki-tfidf

Web

tf-idf model

web-tfidf

X: tf-idf (10k top frequent words in Wikipedia)

input : x

Relu

L1 (10K, 1K)

Relu

L2 (1K, 1K)

L31 (1k, 16)

L32 (1k, 16)

sigmoid

linear

S

continuous latent vector s

x'

Loss = KL(N(0,1) | N(μ,σ)) + binary_crossentropy(x, x')

Encoder

Decoder

Milad Alshomary

Pipeline for TR extraction

05.07.2018

82
- Performance of candidacy functions on different thresholds:

Thresholds between (1 to 1000 and step of 5)

Documents from sample who have number of aligned docs <= 10

Documents from sample who have number of aligned docs > 10
- Performance of candidacy functions on different thresholds:

Thresholds between (1 to 1000 and step of 5)

![Graphs showing precision and recall for different thresholds](image)

Documents from sample who have number of aligned docs <= 10

Documents from sample who have number of aligned docs > 10
- **Candidate Elimination procedure over the cluster:**

![Diagram of Candidate Elimination procedure]

1. **Load hdfs file**
2. **Partition by doc_idx into 100 partitions**
3. **Broadcast object**
   - Partition x of tfidf-rdd1 is accessible to all partitions of tfidf-rdd2
4. **On parallel compute pairwise scoring on each partition with the broadcasted object**
5. **Persist result into hdfs**

- **ID1_1**
  - (ID2_1, score1)
  - (ID2_7, score2)
  - (ID2_5, score3)

- **ID1_2**
  - (ID2_2, score1)
  - (ID2_5, score2)
  - (ID2_4, score3)
  - (ID2_8, score4)

- **ID1_n**
  - (ID2_3, score1)

- **ID1_x ∈ D1 Dataset**
- **ID2_x ∈ D2 Dataset**
- **score ∈ [0, 1]**
- **score<i> > score<i+1>** (List of candidates are sorted by score)
- Hash based Candidate Elimination procedure over the cluster:

1. Create inverted index
2. Join two inverted index
3. Filter out document that does not exist in the shared index
4. Broadcast the shared index and wiki tfidfs
5. Partition
6. On parallel compute pairwise scoring on each partition with the broadcasted objects
Hash based Candidate Elimination procedure over the cluster:

1. Create inverted index
2. Join two inverted index
3. Filter out document that does not exist in the shared index
4. Broadcast the shared index and wiki tfidfs
5. Broadcast object 2
6. On parallel compute pairwise scoring on each partition with the broadcasted objects
- **Heuristics:**
  - \( H1: \text{ne}_\text{sim} \in (0.5, 1.0] \text{ AND } \text{10grams}_\text{sim} > 0.5 \text{ AND } (\text{s}_\text{percent}_\text{reused} < 0.5 \text{ or } \text{t}_\text{percent}_\text{reused} < 0.5) \Rightarrow \textbf{content reuse} \text{ otherwise } \textbf{structure reuse} \)
  - 6700 content reuse cases only
  - Validation on two random samples of size 100:
    - Sample 1:
      | Structure reuse | Content reuse |
      |-----------------|--------------|
      | 100%            | 58%          |
    - Sample 2 (Text 1 or Text 2 > 200):
      | Structure reuse | Content reuse |
      |-----------------|--------------|
      | 100%            | 73%          |