Word Embeddings in Search Engines, Quality Evaluation
Neural Networks are widely used with high rate of success. But can we reproduce those results in IR?
Motivation

State of the art for query processing: BM25
Not sufficient for synonym detection or combined words.

- Current solution:
  Fetch additional personal data.
Motivation

Introducing Neural Networks into Text similarity.

Lots of work recently:
- Mikolov et al. (2013)
- Mikolov et al. (2014)
- Kusner et al. (2015)

Use those results in Information Retrieval.
System components

- Algorithms
  - Best Match 25
  - Word Vectors
    - Document Vectors
  - Word Mover‘s Distance

- Datasets
Algorithms

Best Match 25

• Tries to solve problems that basic TF-IDF systems have.

BM25 = TF*/ α

Where

• $TF^* = \frac{TF(k + 1)}{k + TF}$ where $k \in [0, \infty)$

• $\alpha = (1 - b) + b \times \frac{DL}{AVDL}$
Algorithms

Word Vectors

- Mikolov et al. (2013)
- Uses shallow RNN to train
- Gives a point in space to each word
- Similarities can be computed
Algorithms

Document Vectors

- Mikolov et al. (2014)
- Uses Word vectors.
- Document to document similarity.

But:
- Training expensive
Word Mover’s Distance

- Another approach to doc to doc similarity
- Minimizes cumulative distance between words.

But:
- Processing expensive

![Diagram illustrating Word Mover's Distance with word2vec embedding]
Datasets

- WMT 2011 News Crawl data.
- Movies dataset.
- NPL dataset.
Datasets

WMT 2011 News Crawl data.

- 1 Billion words dataset.
- Cleaned and divided in training and testing.
- Useful for training models.
Datasets

Movies dataset.

- Moderate size of 140K documents.
- Relatively cleaned, but has some non english.
- With benchmark for testing.
Datasets

NPL dataset.

- Small dataset of 11K documents.
- Known for giving bad results in benchmarking.
- 93 queries for testing.
Complete flowchart

- Query
- Word Vector
- Document Centroid Vector Matrix
- Index
- BM25 scored Vector Space Matrix
- Final Ranking
- BM25 Ranking
Tried techniques

- Document vectors
- WMD ranking
  - WMD reordering
- Synonyms generation
- Word Averages
  - Word averages combined with BM25
    - Optimization
Document vectors

- Fast computation (matrix operations available)
  But:
  - Extremely expensive to compute.
  - Small corpus → Bad embeddings
  - Horrible results.
Word Mover‘s Distance

• Extremely slow to compute.
• Query and Document semantic spaces differ.
• Results made some sense, but were not good.
Word Mover‘s Distance

- Extremely slow to compute.
- Query and Document semantic spaces differ.
- Results made some sense, but were not good.

WMD reordering:
- Solved slowness
- Worsens BM25 results
Generating synonyms

- Tried to extend the query.

But:

- Words have more than one meaning.
- Most added words did not have the correct meaning.
- Others were combined words and couldn’t be used.
Word Averages

- A document is a cluster of words.
- With word embeddings, compute the centroid.
- Distance between clusters is a similarity measure
Word Averages

However, long documents are bad with this.

- Combine this ranking with BM25.
- Merge them with according to an alpha in [0, 1]
Optimization

Word Averages technique had hyperparameters optimized.

Optimization: local search.

- MAP optimization in Movies.
- Best of three out of MP@3, MP@R and MAP with mixed databases.
Results

- Most techniques were mixed with BM25.
- Some bad results are also shown.
# Movies dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Hyper-parameters</th>
<th>MP@3</th>
<th>MP@R</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base mixed &amp; mean vectors (db2-opt)</td>
<td>k=1.4 b=0.6822</td>
<td>0.5667</td>
<td>0.3639</td>
<td>0.3641</td>
</tr>
<tr>
<td></td>
<td>alpha=0.6695</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base mixed &amp; mean vectors (base)</td>
<td>k=1.75 b=0.7</td>
<td>0.5</td>
<td>0.2881</td>
<td>0.3162</td>
</tr>
<tr>
<td></td>
<td>alpha=0.67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base BM25</td>
<td>k=1.75 b=0.7</td>
<td>0.4333</td>
<td>0.2869</td>
<td>0.3092</td>
</tr>
<tr>
<td>Base mixed &amp; mean vectors (mixed-opt)</td>
<td>k=6.987 b=0.3199</td>
<td>0.5</td>
<td>0.3105</td>
<td>0.2964</td>
</tr>
<tr>
<td></td>
<td>alpha=0.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc2vec mixed &amp; base</td>
<td>k=1.75 b=0.7</td>
<td>0.4333</td>
<td>0.2869</td>
<td>0.3072</td>
</tr>
<tr>
<td></td>
<td>alpha=0.67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BM25 with WMD reorder</td>
<td>k=1.75 b=0.7</td>
<td>0.3333</td>
<td>0.2423</td>
<td>0.2914</td>
</tr>
</tbody>
</table>

| | | | | |
## NPL dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Hyper-parameters</th>
<th>MP@3</th>
<th>MP@R</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base mixed &amp; mean vectors (base)</td>
<td>k=1.75 b=0.7 alpha=0.67</td>
<td>0.4265</td>
<td>0.2210</td>
<td>0.1850</td>
</tr>
<tr>
<td>Base BM25</td>
<td>k=1.75 b=0.7</td>
<td>0.4194</td>
<td>0.2375</td>
<td>0.2095</td>
</tr>
<tr>
<td>Base mixed &amp; mean vectors (mixed-opt)</td>
<td>k=6.987 b=0.3199 alpha=0.4396</td>
<td>0.4121</td>
<td>0.2326</td>
<td>0.1904</td>
</tr>
<tr>
<td>Doc2vec mixed &amp; base</td>
<td>k=1.75 b=0.7 alpha=0.67</td>
<td>0.4194</td>
<td>0.2375</td>
<td>0.2095</td>
</tr>
<tr>
<td>BM25 with WMD reoder</td>
<td>k=1.75 b=0.7</td>
<td>0.2115</td>
<td>0.1940</td>
<td>0.1543</td>
</tr>
</tbody>
</table>
Improvements

- Fix the method combining function.
- Change functions to work with cosine similarity.
- Thoroughly test the engine with a larger database.
- Build a new Word Vector Model that includes the index.
Future Work

- Get computer power to build a quality Document Vector model.
- Work into dividing multi themed document into single themed subdocuments.
- Different ways to combine the Word Model method with the base method.
Thank you!

Any question?
Matrix the movie

[['Service_NASDAQ_MTRX', 0.6663896441459656),
 ('Orecel_R_Cellular', 0.5899302959442139),
 ('NASDAQ_MTRX', 0.5458115935325623),
 ('System_Automotive_Finishes', 0.5215746164321899),
 ('Nasdaq_MTRX', 0.5153658390045166),
 ('Neo_Keanu_Reeves', 0.5114843845367432),
 ('AMOLED_Active', 0.5107872486114502),
 ('Requirements_Validation', 0.5089198350906372),
 ('Co._MTRX', 0.49655118584632874),
 ('Neo', 0.4950855672359466)],
[['this', 0.5937378406524658),
 ('in', 0.5429296493530273),
 ('that', 0.5262569785118103),
 ('ofthe', 0.5150282382965088),
 ('another', 0.47483527660369873),
 ('however', 0.4748331904411316),
 ('one', 0.4665869176387787),
 ('entire', 0.4619824290275574),
 ('its', 0.4605940580368042),
 ('which', 0.4595310688018799)],
[['film', 0.8676770329475403),
 ('movies', 0.8013108968734741),
 ('films', 0.7363011837005615),
 ('move', 0.6830361485481262),
 ('Movie', 0.6693680286407471),
 ('horror_flick', 0.6577848196029663),
 ('sequel', 0.657779335975647),
 ('Guy_Ritchie_Revolver', 0.650975227355957),
 ('romantic_comedy', 0.6413198709487915),
 ('flick', 0.6321909427642822)]
Type your query:

>>> Matrix the movie
Score: 28.729940520400532, doc preview:The Matrix Revisited
Score: 27.40164185968183, doc preview:Cymasonics - Matrix Optimizer 1.0.1
Score: 25.55098586267934, doc preview:Matrix Revisited
Score: 25.293376140108133, doc preview:Escape the Matrix
Score: 25.281932159196277, doc preview:The Matrix Revolutions
Score: 24.97942046004182, doc preview:The Matrix Reloaded
Score: 24.736015248150533, doc preview:Dot Matrix
Score: 24.26170579128672, doc preview:Bigger Questions: The Psychic Matrix
Score: 21.11061410978737, doc preview:MTV Movie Awards Reloaded MTV Movie Awards
Score: 20.228968126579062, doc preview:The Matrix

Type your query:
Document Vectors

Type your query:

>>> Matrix the movie
Score: 0.9870679753510716, doc preview:The Vagabond King  The Vagabond King is a
Score: 0.987014094675869, doc preview:Hughie Green, Most Sincerely  Drama about
Score: 0.9870100482156816, doc preview:The Mystery Train  The Mystery Train is a
Score: 0.98698864866112621, doc preview:The West Wittering Affair  When Jamie acc
Score: 0.986957097939487, doc preview:Goopy Geer  Goopy Geer is a 1932 Merrie M
Score: 0.9869189952375426, doc preview:South Wind  South Wind is a 2012 short, a
Score: 0.9869139825363086, doc preview:Vipers in the Grass  Vipers in the Grass
Score: 0.9869126445546323, doc preview:Mina... fuori la guardia Mina... fuori l
Score: 0.9868904781278359, doc preview:Every 9 Seconds  Every 9 Seconds, is a 19
Score: 0.986889227525191, doc preview:Double Suicide  Double Suicide is a 1969

Type your query:

>>>
Average Vectors

Type your query:

>>> Matrix the movie
Score: 0.5281055031614802, doc preview: Matrix Matrix is a 1999 short film direc
Score: 0.5021051846332215, doc preview: Ressha daikosin the movie Ressha number-
Score: 0.49429904516514363, doc preview: 976-Evil II 976-EVIL II, also known as 9
Score: 0.4888701910441772, doc preview: The Matrix Revolutions The Matrix Revolu
Score: 0.48657254848093734, doc preview: Aakramana Aakramana is a horror Kannada m
Score: 0.4851319059045448, doc preview: Dot Matrix Dot Matrix is a 2013 short fi
Score: 0.48273688792727787, doc preview: MTV Movie Awards Reloaded MTV Movie Awar
Score: 0.4781561421080653, doc preview: Avatar Avatar, also known as Matrix Hunt
Score: 0.477699585128027, doc preview: Nagahen: Nagina Part II Nagahen: Nagina
Score: 0.4773364072848646, doc preview: Action Man: X Missions – The Movie Actio
Type your query:

>>> 


BM25 + Doc2vec

Type your query:

> >>> Matrix the movie
Score: 0.9996698308981523, doc preview:The Matrix Revisited The Matrix Revisited
Score: 0.968516293254379, doc preview:Cymasonics - Matrix Optimizer 1.0.1 Cyma
Score: 0.9254327746671083, doc preview:Matrix Matrix is a 1999 short film direc
Score: 0.919475721637181, doc preview:Escape the Matrix Escape the Matrix is a
Score: 0.919196938916635, doc preview:The Matrix Revolutions The Matrix Revolu
Score: 0.9122558065757396, doc preview:The Matrix Reloaded The Matrix Reloaded
Score: 0.9065688353398629, doc preview:Dot Matrix Dot Matrix is a 2013 short fi
Score: 0.8954855058983764, doc preview:Bigger Questions: The Psychic Matrix Big
Score: 0.8220502203209992, doc preview:MTV Movie Awards Reloaded MTV Movie Awar
Score: 0.8016736694238987, doc preview:The Matrix The Matrix is a 1999 American

Type your query:

> >>>
BM25 + Avg vectors

Type your query:

>>> Matrix the movie
Score: 0.9568451994018551, doc preview:The Matrix Revisited The Matrix Revisite
Score: 0.925920134020497, doc preview:Matrix Matrix is a 1999 short film direc
Score: 0.8950959991781918, doc preview:The Matrix Revolutions The Matrix Revolu
Score: 0.8667920845695645, doc preview:Escape the Matrix Escape the Matrix is a
Score: 0.864615677961966, doc preview:The Matrix Reloaded The Matrix Reloaded
Score: 0.848256950334392, doc preview:Bigger Questions: The Psychic Matrix Big
Score: 0.7940525195308585, doc preview:MTV Movie Awards Reloaded MTV Movie Awar
Score: 0.7479453341396294, doc preview:The Matrix The Matrix is a 1999 American
Score: 0.7357574976003911, doc preview:Avatar Avatar, also known as Matrix Hunt
Score: 0.734482111213234, doc preview:The Animatrix The Animatrix is a 2003 Am
BM25 with reordering

Type your query:

>>> Matrix the movie
Score: 0.448671799813948, doc preview: Computer Boy  Computer Boy is a 2000 short
Score: 0.4313591792616431, doc preview: Dot Matrix  Dot Matrix is a 2013 short fi
Score: 0.42416095616214616, doc preview: The Matrix  The Matrix is a 1999 American
Score: 0.4228804841131871, doc preview: The Animatrix  The Animatrix is a 2003 Am
Score: 0.42258338537458867, doc preview: MTV Movie Awards Reloaded  MTV Movie Awar
Score: 0.4176156392958816, doc preview: The Bloody Fists The Bloody Fists, aka D
Score: 0.40708154671318164, doc preview: Scary Movie  Scary Movie is a 2000 horror
Score: 0.40505716013508397, doc preview: Armitage III  Armitage III is a 1995 cybe
Score: 0.3954435658175758, doc preview: The Matrix Revisited  The Matrix Revisite
Score: 0.39521036307799345, doc preview: Scary Movie 3  Scary Movie 3 is a 2003 Am

Type your query:

>>>