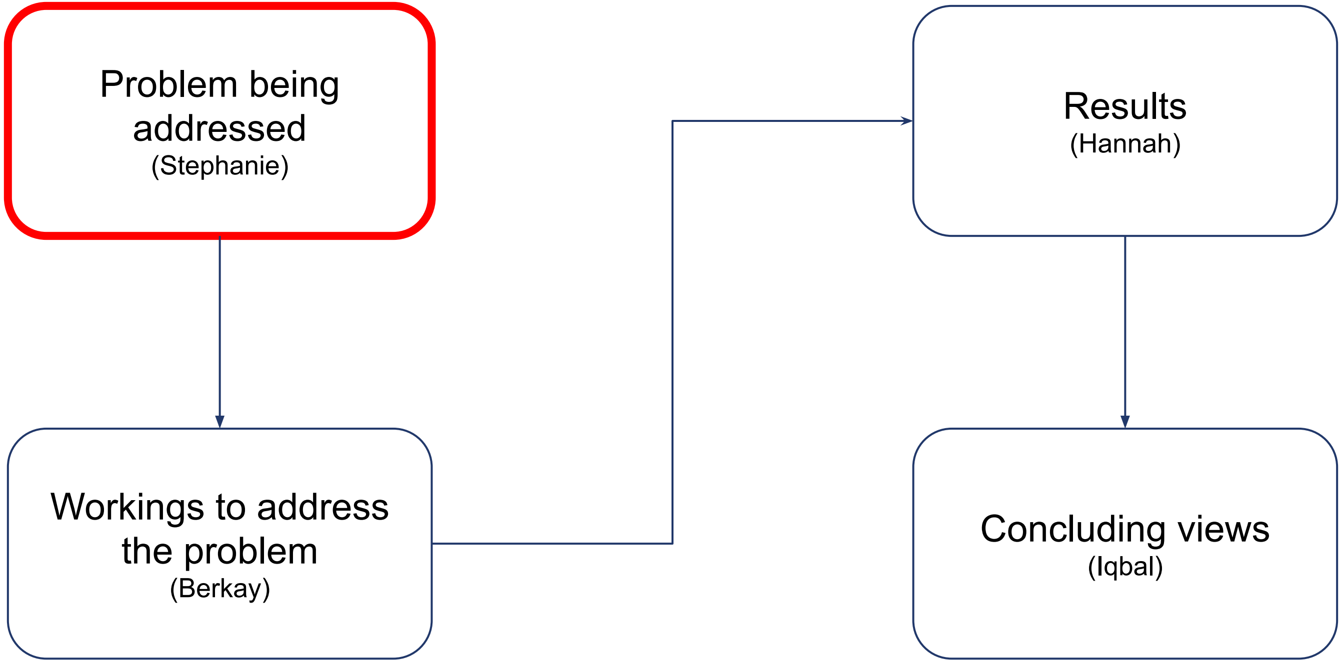
# Improving Document Ranking with Dual Word Embeddings

ECS735P Information Retrieval - Group 18

Stephanie Nicole Garibay Lim
Berkay Dur
Hannah Melkemaryam Claus
Iqbal Singh

# Presentation Structure



# Improving Document Ranking with Dual Word Embeddings

## Problems with information retrieval:

1. Document and query may contain different vocabulary
2. Irrelevant document containing terms from the query
3. Linking query with document
4. Incorporation of new document content

## Goal:

To model the relationship between the query and the document content, without memorised click data

(Nalisnick *et al.*, 2016)

# Solutions to the Goal

## Possible Solution: TF-IDF BM25

(1) Count repetition of query terms in the document
(2) Original query terms
(3) Poisson model

## Improving the Solution

(a) use word occurrences as evidence of relevance
(b) Do not count term repetition
(c) Consider the relationship between the terms (query and document)

## Proposal: Dual Embedding Space Model (DESM)

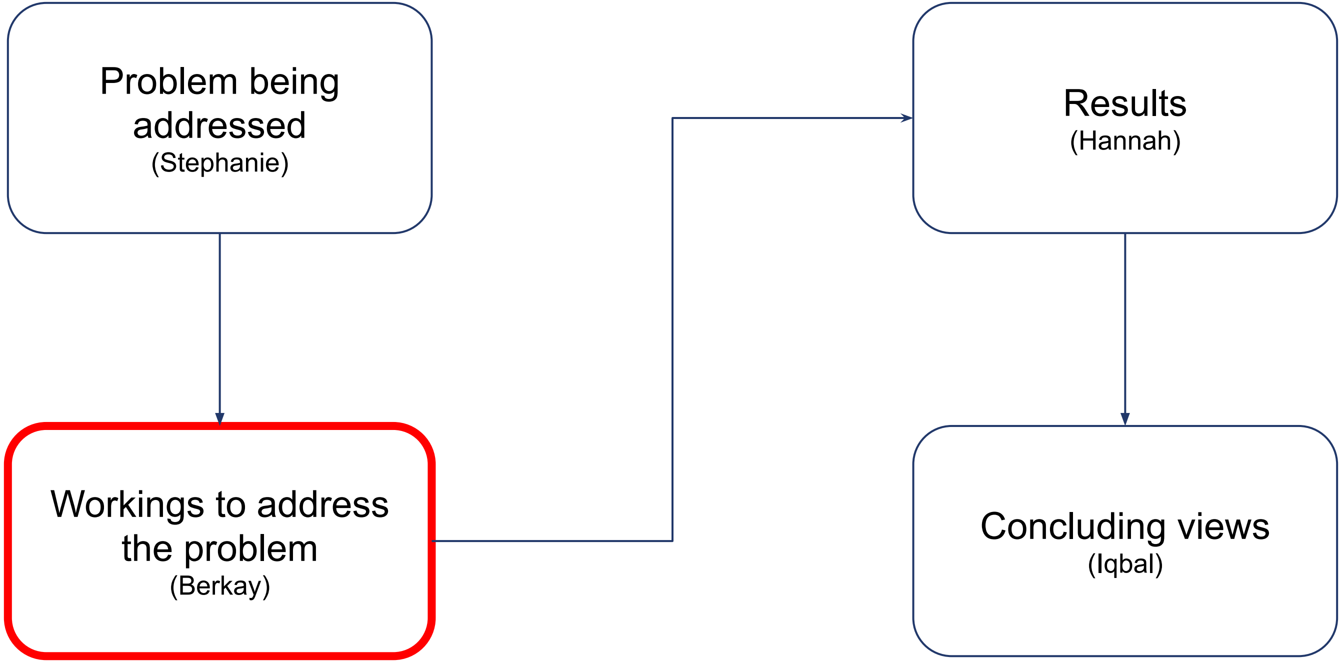
### Word2Vec model

- Input and Output embedding spaces

### DESM Model

- Query words
- Document words

# Presentation Structure



# Workings - Overall workings of the DESM

**Dual Embedding Space Model**

Train Word2Vec and use the embedding representation for Q and D.

D - Use embeddings for document body
Q - Use embeddings for query corpus

**Ranking Function**

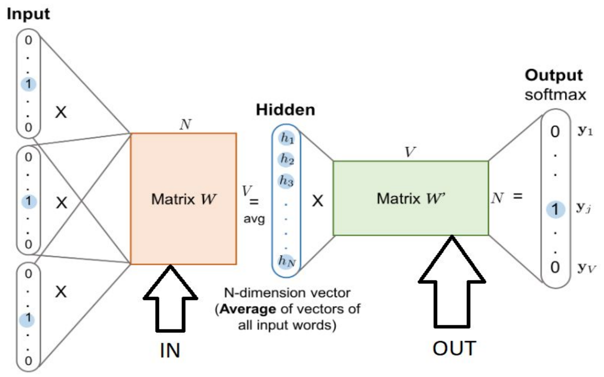
$$DESM(Q, D) = \frac{1}{|Q|} \sum_{q_i \in Q} \frac{\mathbf{q}_i^T \overline{\mathbf{D}}}{\|\mathbf{q}_i\| \|\overline{\mathbf{D}}\|}$$

Where:

$$\overline{\mathbf{D}} = \frac{1}{|D|} \sum_{\mathbf{d}_j \in D} \frac{\mathbf{d}_j}{\|\mathbf{d}_j\|}$$

**Which Embedding?**

Which Embeddings should Q and D representations come from?




There are 4 possible combinations:

1. Q - IN, D - IN
2. Q - IN, D - OUT
3. Q - OUT, D - IN
4. Q - OUT, D - OUT

$DESM_{IN-IN}(Q, D)$
$DESM_{IN-OUT}(Q, D)$
$DESM_{OUT-IN}(Q, D)$
$DESM_{OUT-OUT}(Q, D)$

(Weng, 2017, modified; Mitra et al., 2016)

# Workings - Overall workings of the DESM

**Ranking Functions of Paper:**

$$DESM_{IN-OUT}(Q,D) = \frac{1}{|Q|}\sum_{q_i \in Q} \frac{q_{IN,i}^T \overline{D_{OUT}}}{\|q_{IN,i}\| \|\overline{D_{OUT}}\|}$$

$$DESM_{IN-IN}(Q,D) = \frac{1}{|Q|}\sum_{q_i \in Q} \frac{q_{IN,i}^T \overline{D_{IN}}}{\|q_{IN,i}\| \|\overline{D_{IN}}\|}$$

| | seahawks | |
|---|---|---|
| IN-IN | OUT-OUT | IN-OUT |
| seahawks | seahawks | seahawks |
| 49ers | broncos | highlights |
| broncos | 49ers | jerseys |
| packers | nfl | tshirts |
| nfl | packers | seattle |
| steelers | steelers | hats |

**Why do we get different results?**

IN-OUT cosine similarity is higher for words that co-occur often in the training corpus and are **topically** similar.

IN-IN cosine similarity is higher for words that are **typically** similar.

(Mitra et al., 2016)

# Workings - Overall workings of the DESM

*Albuquerque is the most populous city in the U.S. state of New Mexico. The high-altitude city serves as the county seat of Bernalillo County, and it is situated in the central part of the state, straddling the Rio Grande. The city population is 557,169 as of the July 1, 2014, population estimate from the United States Census Bureau, and ranks as the 32nd-largest city in the U.S. The Metropolitan Statistical Area (or MSA) has a population of 902,797 according to the United States Census Bureau's most recently available estimate for July 1, 2013.*

*Allen suggested that they could program a BASIC interpreter for the device; after a call from Gates claiming to have a working interpreter, MITS requested a demonstration. Since they didn't actually have one, Allen worked on a simulator for the Altair while Gates developed the interpreter. Although they developed the interpreter on a simulator and not the actual device, the interpreter worked flawlessly when they demonstrated the interpreter to MITS in Albuquerque, New Mexico in March 1975; MITS agreed to distribute it, marketing it as Altair BASIC.*

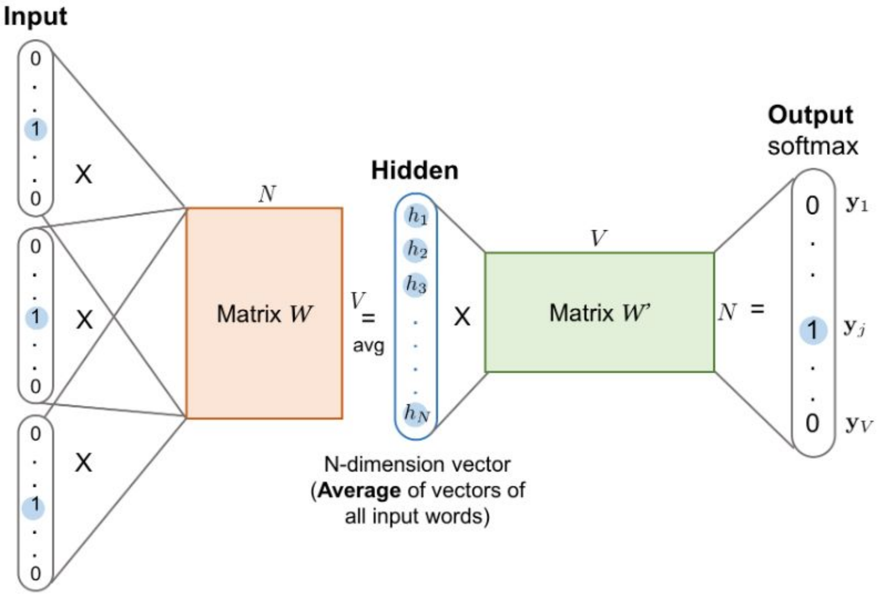
(Mitra et al., 2016)

# Workings - Overall workings of the DESM

**Distributional Semantics**
"You shall know a word by the company it keeps"
- J.R.Firth

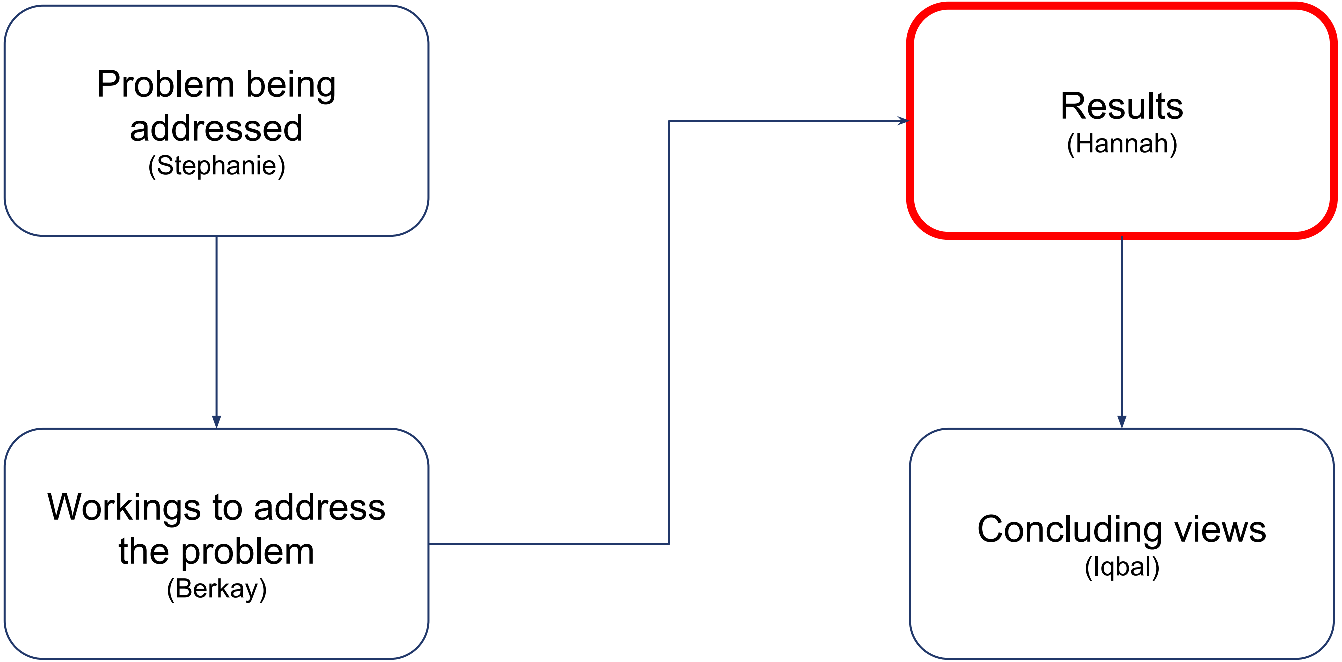
**Word Embeddings**
A dense vector representation of a word that encodes the distributional semantics of a word (I.e., words that have similar meaning will have similar word embeddings).

**Word2Vec**
Neural NLP method where 2 word embeddings are trained.




(Weng, 2017)

# Presentation Structure



# Experiments

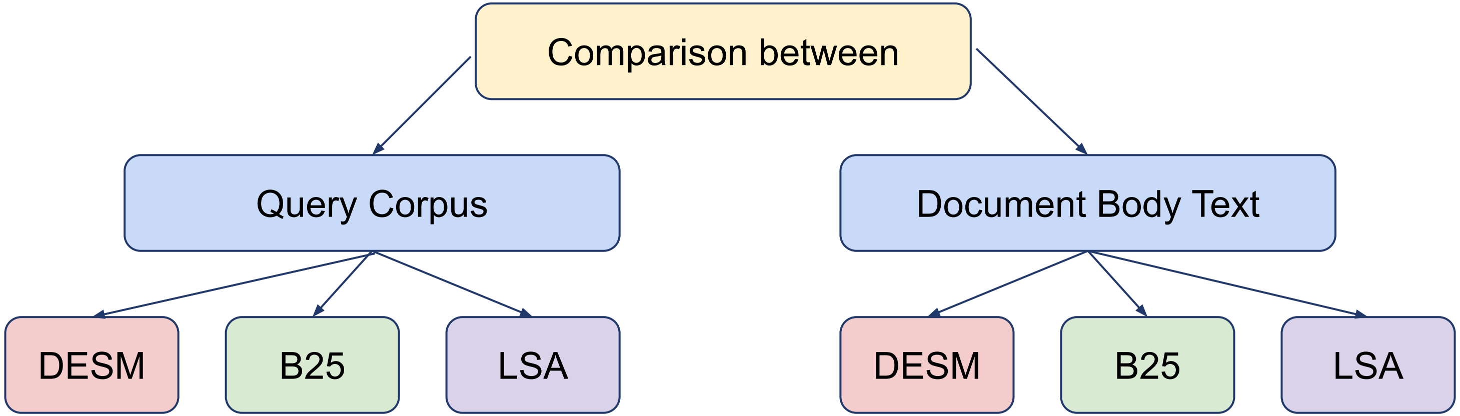
**Models:**

- Dual Embedding Space Model (DESM)
- BM25, a traditional count-based method
- Latent Semantic Analysis (LSA), a traditional vector-based method

# Experiments

**Dataset:**

- Sampled from Bing's large scale query logs
- Body text for all the candidate documents are extracted from Bing's document index



# Experiments

**Performance Measure:**

- Normalized discounted cumulative gain (NDCG)

➔ is a measure of ranking quality that evaluates the effectiveness of an information retrieval system

➔ takes into account the relevance of each document to the query and the rank position of the document in the search results

(Järvelin and Kekäläinen, 2017)

# Findings on Query Corpus

Table 3: NDCG results comparing the $DESM_{IN-OUT}$ with the BM25 and the LSA baselines. The $DESM_{IN-OUT}$ performs significantly better than both the BM25 and the LSA baselines at all rank positions. It also performs better than the $DESM_{IN-IN}$ on both the evaluation sets. The DESMs using embeddings trained on the query corpus also performs better than if trained on document body text. The highest NDCG values for every column is highlighted in bold and all the statistically significant ($p < 0.05$) differences over the BM25 baseline are marked with the asterisk (*).

| | Explicitly Judged Test Set | | | Implicit Feedback based Test Set | | |
|---|---|---|---|---|---|---|
| | NDCG@1 | NDCG@3 | NDCG@10 | NDCG@1 | NDCG@3 | NDCG@10 |
| BM25 | 23.69 | 29.14 | 44.77 | 13.65 | 27.41 | 49.26 |
| LSA | 22.41* | 28.25* | 44.24* | 16.35* | 31.75* | 52.05* |
| DESM (IN-IN, trained on body text) | 23.59 | 29.59 | 45.51* | 18.62* | 33.80* | 53.32* |
| DESM (IN-IN, trained on queries) | 23.75 | 29.72 | 46.36* | 18.37* | 35.18* | 54.20* |
| DESM (IN-OUT, trained on body text) | 24.06 | 30.32* | 46.57* | 19.67* | 35.53* | 54.13* |
| DESM (IN-OUT, trained on queries) | **25.02*** | **31.14*** | **47.89*** | **20.66*** | **37.34*** | **55.84*** |

(Mitra et al., 2016)

# Findings on Document Body Text

**Table 4: Results of NDCG evaluations under the non-telescoping settings. Both the DESM and the LSA models perform poorly in the presence of random irrelevant documents in the candidate set. The mixture of $DESM_{IN-OUT}$ with BM25 achieves the best NDCG. The best NDCG values are highlighted per column in bold and all the statistically significant ($p < 0.05$) differences with the BM25 baseline are indicated by the asterisk (*)**

| | Explicitly Judged Test Set | | | Implicit Feedback based Test Set | | |
|---|---|---|---|---|---|---|
| | NDCG@1 | NDCG@3 | NDCG@10 | NDCG@1 | NDCG@3 | NDCG@10 |
| BM25 | 21.44 | 26.09 | 37.53 | 11.68 | 22.14 | 33.19 |
| LSA | 04.61* | 04.63* | 04.83* | 01.97* | 03.24* | 04.54* |
| DESM (IN-IN, trained on body text) | 06.69* | 06.80* | 07.39* | 03.39* | 05.09* | 07.13* |
| DESM (IN-IN, trained on queries) | 05.56* | 05.59* | 06.03* | 02.62* | 04.06* | 05.92* |
| DESM (IN-OUT, trained on body text) | 01.01* | 01.16* | 01.58* | 00.78* | 01.12* | 02.07* |
| DESM (IN-OUT, trained on queries) | 00.62* | 00.58* | 00.81* | 00.29* | 00.39* | 01.36* |
| BM25 + DESM (IN-IN, trained on body text) | 21.53 | 26.16 | 37.48 | 11.96 | 22.58* | 33.70* |
| BM25 + DESM (IN-IN, trained on queries) | **21.58** | 26.20 | 37.62 | 11.91 | 22.47* | 33.72* |
| BM25 + DESM (IN-OUT, trained on body text) | 21.47 | 26.18 | 37.55 | 11.83 | 22.42* | 33.60* |
| BM25 + DESM (IN-OUT, trained on queries) | 21.54 | **26.42*** | **37.86*** | **12.22*** | **22.96*** | **34.11*** |

(Mitra et al., 2016)

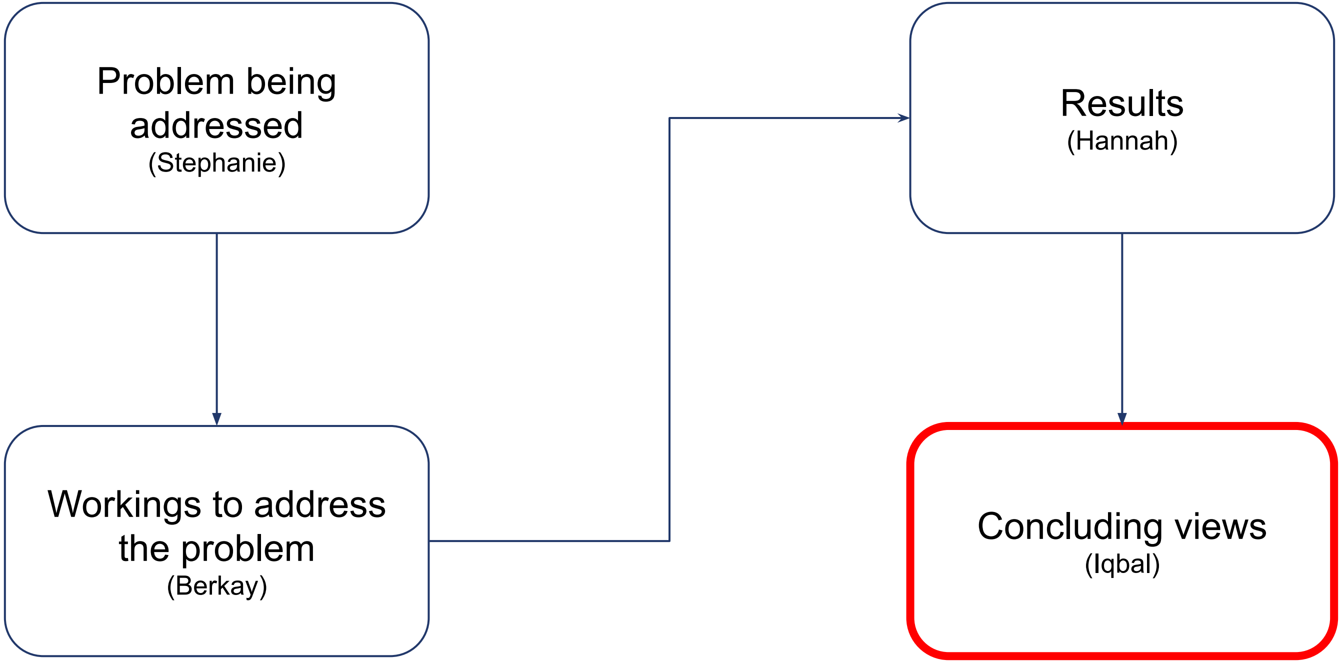
# Findings

- On Query Corpus, the DESM IN−OUT performs best
- The **mixture** of DESM IN−OUT (trained on queries) and BM25 gives the best NDCG result under the non-telescoping settings

➔ Hence, DESM IN−OUT is the most discriminating feature for the relevant and the irrelevant documents retrieved by a first stage retrieval system

➔ the DESM is primarily suited for ranking at top positions or in conjunction with other document ranking features

# Presentation Structure



# Conclusion and Further expansion of research

## Potential contributions

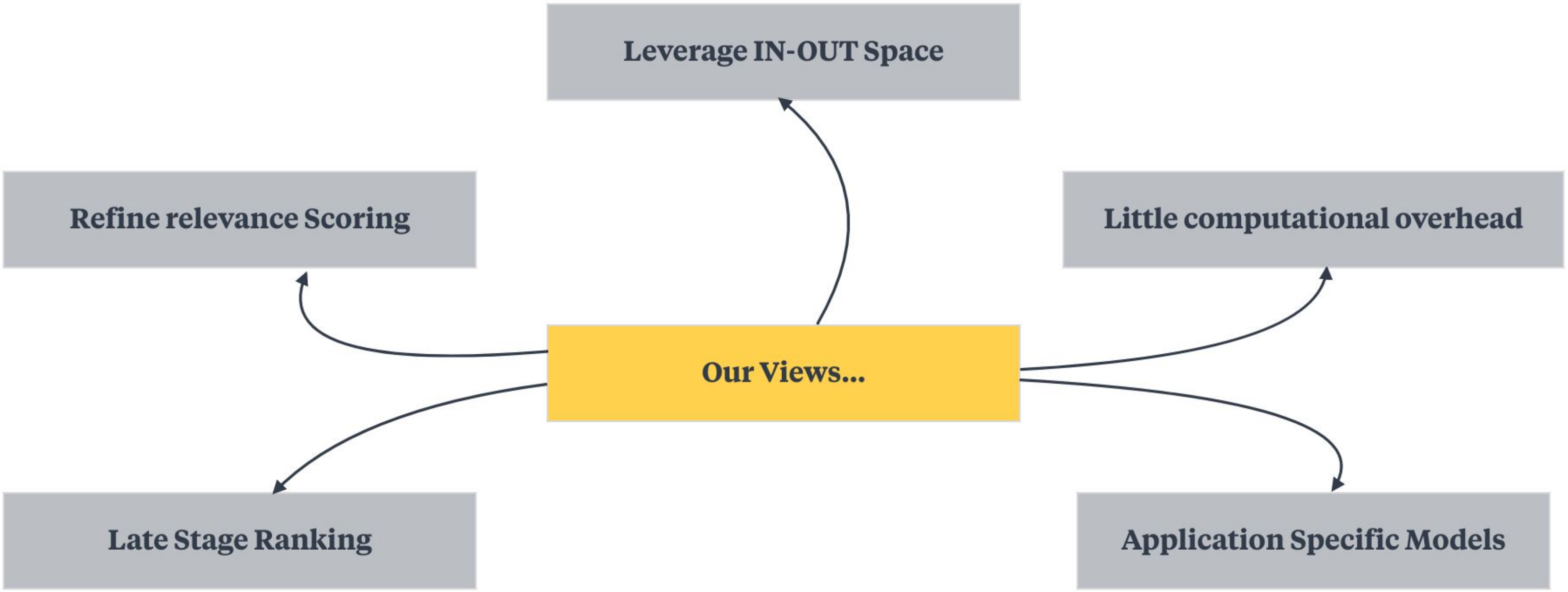
- All words in a document to contribute to rankings
- Ranking attuned to **semantic subtleties**
- Leveraging, output embeddings
- Refining relevance scoring
- Incorporated into other stages of the IR pipeline

## Benefits of IN-OUT space

- Ranking via proximity in IN-OUT space is better for retrieval than IN-IN based rankers
- Findings emphasis that usage of the CBOW and SG models is **application dependent**

## Future directions

- Incorporating Out of Vocabulary words
- Document Length Normalisation
- IN-IN and IN-OUT based distances into other stages of the IR pipeline.
- Incorporating pre trained embeddings (Glove, ELMo, BERT)


Leverage IN-OUT Space
Refine relevance Scoring
Little computational overhead
Our Views...
Late Stage Ranking
Application Specific Models

# Appendix

**Presentation's Journal Paper:**

Nalisnick, E. *et al.* (2016) "Improving document ranking with dual word embeddings," *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, pp. 83–84. Available at: https://doi.org/10.1145/2872518.2889361.

**Notable citations:**


- Guo, J., Cai, Y., Fan, Y., Sun, F., Zhang, R. and Cheng, X. (2022). Semantic Models for the First-Stage Retrieval: A Comprehensive Review. ACM Transactions on Information Systems, [online] 40(4), pp.1–42. doi:https://doi.org/10.1145/3486250.
- Muhammad, I., Bollegala, D., Coenen, F., Gamble, C., Kearney, A. and Williamson, P. (2021). Document Ranking for Curated Document Databases Using BERT and Knowledge Graph Embeddings: Introducing GRAB-Rank. Big Data Analytics and Knowledge Discovery, pp.116–127. doi:https://doi.org/10.1007/978-3-030-86534-4_10.
- Chy, A.N., Ullah, M.Z. and Aono, M. (2019). Query Expansion for Microblog Retrieval Focusing on an Ensemble of Features. Journal of Information Processing, [online] 27, pp.61–76. doi:https://doi.org/10.2197/ipsjjip.27.61.
- Khattab, O., Hammoud, M. and Elsayed, T. (2020). Finding the Best of Both Worlds. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. doi:https://doi.org/10.1145/3397271.3401076.
- Järvelin, K. and Kekäläinen, J. (2017). IR evaluation methods for retrieving highly relevant documents. ACM SIGIR Forum, 51(2), pp.243–250. doi:https://doi.org/10.1145/3130348.3130374.
- Mitra, B., Nalisnick, E., Craswell, N., Caruana, R. and Redmond, M. (2016). *A Dual Embedding Space Model for Document Ranking*. [online] Available at: https://arxiv.org/pdf/1602.01137.pd
- Weng, L. (2017) *Learning word embedding*, *Lil'Log (Alt + H)*. Available at: https://lilianweng.github.io/posts/2017-10-15-word-embedding/ (Accessed: February 20, 2023).