Maestro: Topic-aware, Intelligent Flashcards

Team 46
Casper Guo
Nick Lovell

Overview

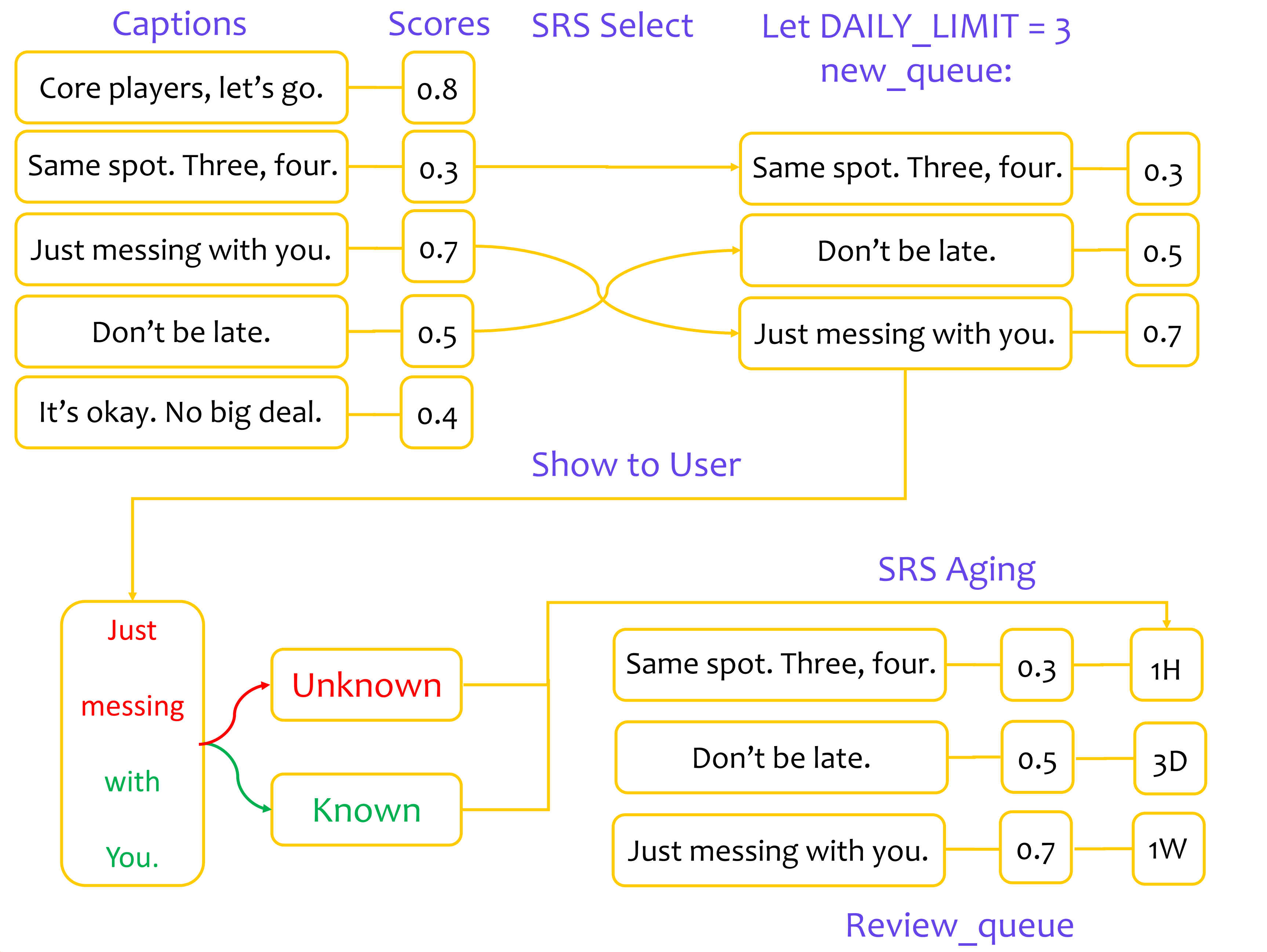
Students often use flashcard apps to supplement foreign language vocabulary learning. Traditional flashcard apps, such as Quizlet, surface user-created cards in random orders. This procedure can be improved in several ways.

1. Spaced Repetition Systems (SRS), such as Anki, schedule card reviews intelligently. If the user reports high confidence about knowing the content of a card, the card will receive a long interstudy interval (ICI), and vice versa. A meta-analysis [1] concludes spaced repetition significantly improves performance compared to massed learning (cramming).

2. Automatic curation of flashcards from a corpus such as the system implemented by Tolmachev et al. [2] optimizes diversity in word senses and sentence constructions and considers sentence difficulty and quality.

Maestro combines the above approaches and additionally elevates user motivation by letting users learn vocabulary via their favorite media. We source our English-Spanish parallel corpus from the OpenSubtitles [3] dataset. For each token in the corpus, we calculate a language importance score based on word frequency and a corpus importance score using topic modelling. The per-token scores are combined into per-sentence scores and are used to rank cards within each review bin calculated by the SRS algorithm.

Spaced Repetition



Captions	Scores
Core players, let's go.	0.8
Same spot. Three, four.	0.3
Just messing with you.	0.7
Don't be late.	0.5
It's okay. No big deal.	0.4

SRS Select

Let DAILY_LIMIT = 3
new_queue:

Same spot. Three, four.	0.3
Don't be late.	0.5
Just messing with you.	0.7

Show to User

Just messing with You.

Unknown

Known

SRS Aging

Same spot. Three, four.	0.3	1H
Don't be late.	0.5	3D
Just messing with you.	0.7	1W

Review_queue

Word Weighting

Preprocessing: We use a spaCy [4] pre-trained pipeline for tokenization and lemmatization and discard all punctuations. Stop words are not removed.

Language Importance: We base the word frequency estimate on the CREA corpus. The words are segmented into four groups that correspond to average vocabulary sizes [5] at each CEFR level and assigned scores between 0 and 1.

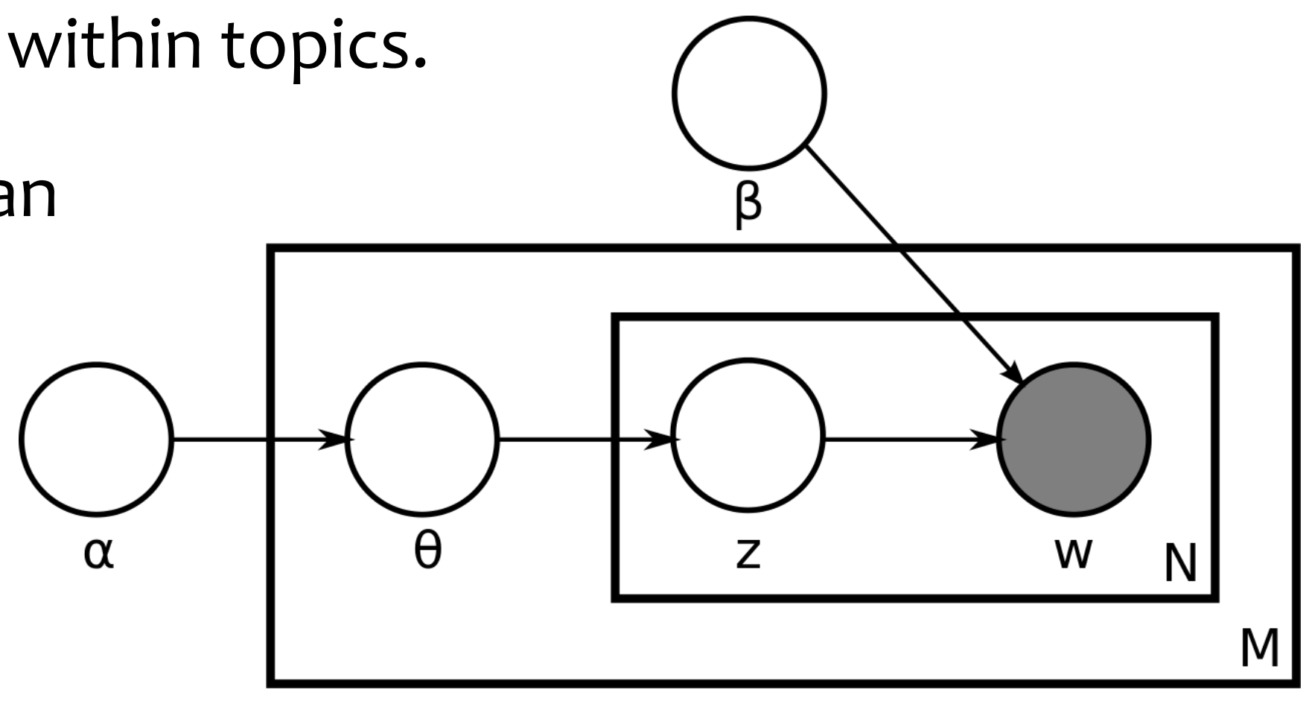
Tf-Idf: We define each line of caption as a document and the entire caption set as the corpus. Idf is computed with add-one smoothing, and we use sublinear tf. The tf-idf scores are min-max normalized to the (0, 1) range.

$$\mathrm{idf} = \log \frac{\#\mathrm{docs} + 1}{\mathrm{df} + 1} + 1$$

$$\mathrm{tf} = 1 + \log \mathrm{tf}$$

Latent Semantic Analysis (LSA): LSA builds upon the tf-idf feature matrix and performs singular value decomposition (SVD) to extract topics based on word distribution patterns. Words are ranked within topics.

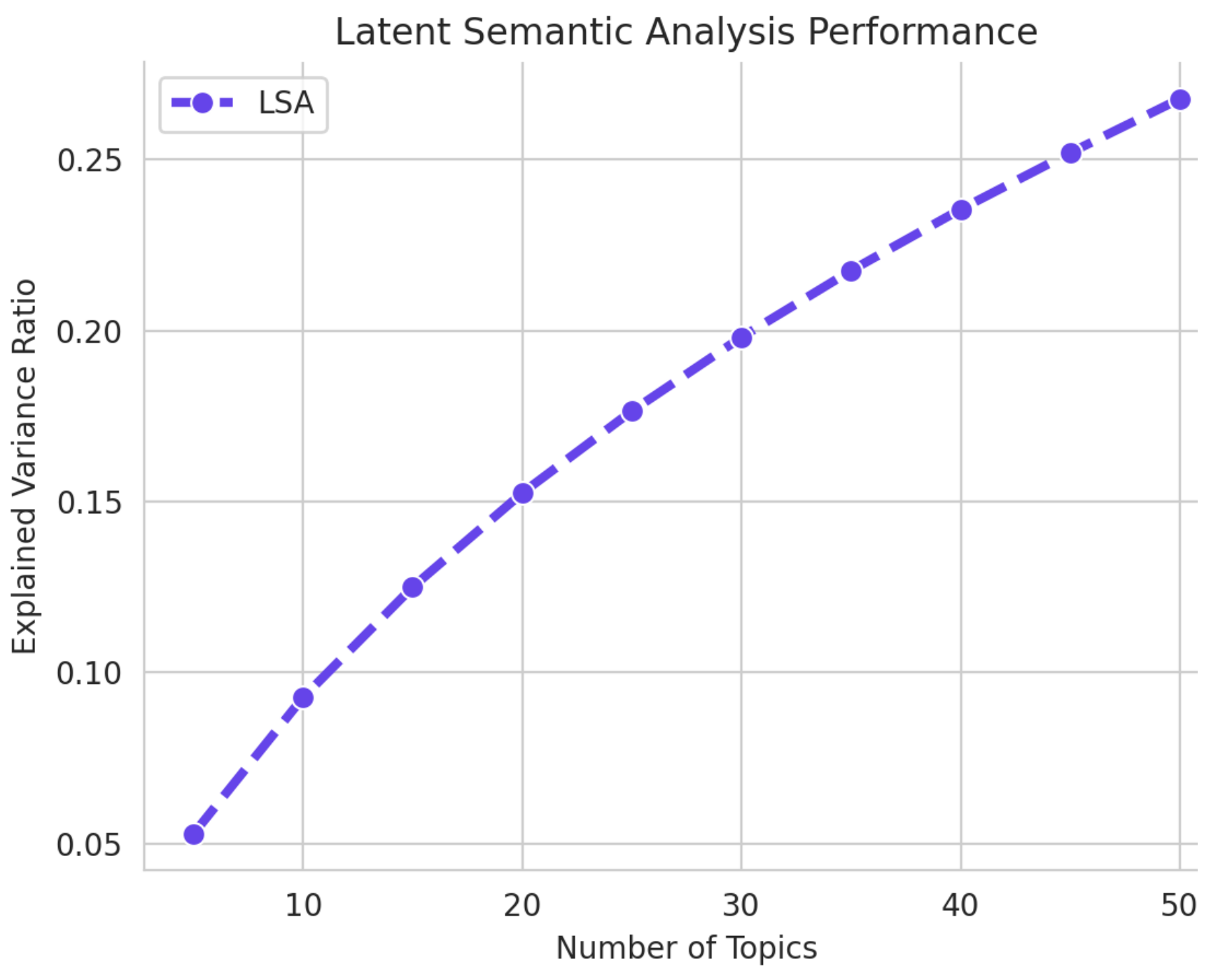
Latent Dirichlet Allocation (LDA): LDA can also be derived from the tf-idf feature matrix. It is a Bayesian method that iteratively optimizes the per-document and per-word topic distributions. Word topics and weights can then be derived.



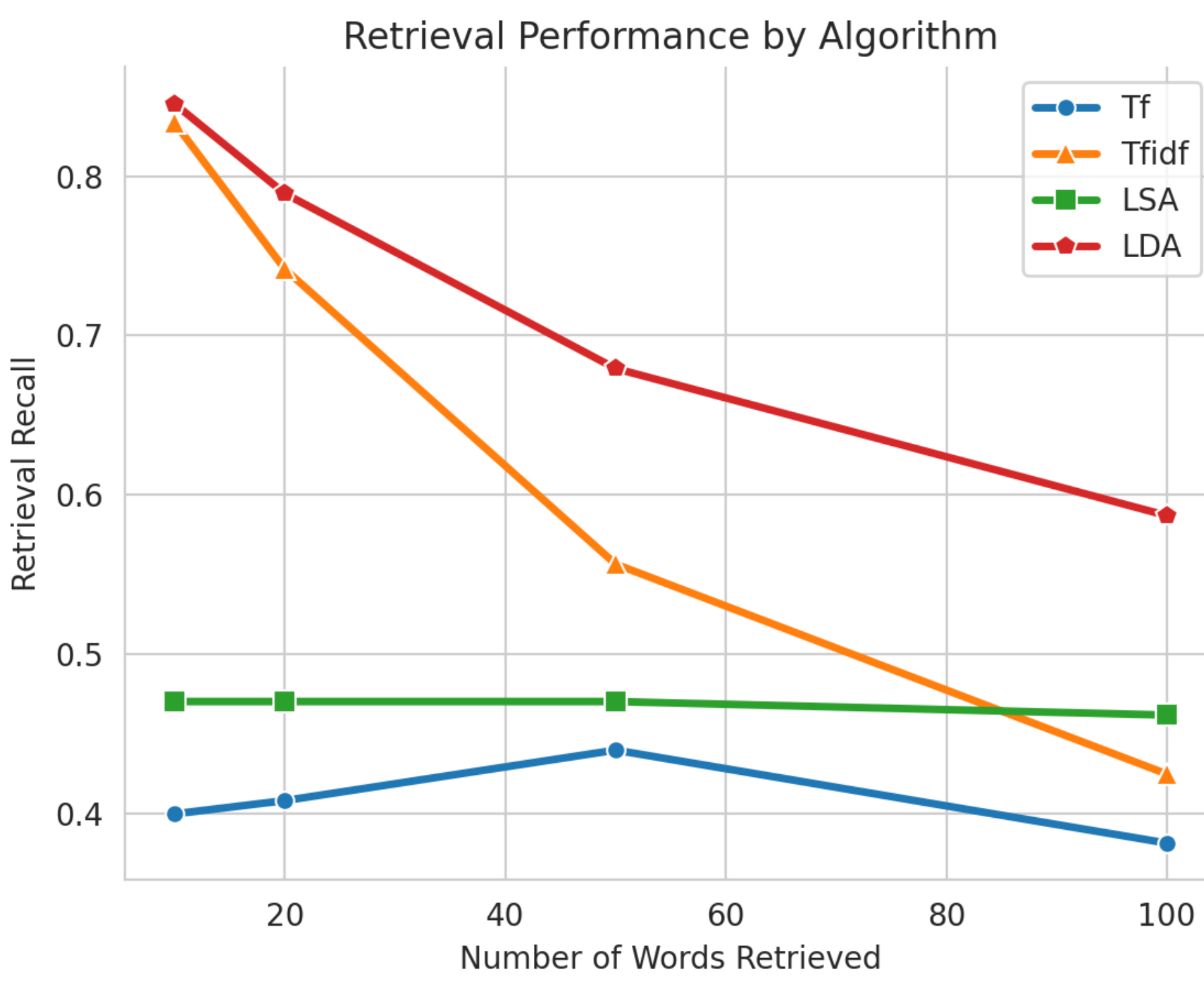
The dependency relationship between LDA random variables

Results

Evaluation Method: We select six movies and retrieve the top words from them with each algorithm. Using the Wikipedia plot summary as the ground truth, the token retrieval recall is calculated. This is motivated by the observation that plot summaries are effectively human labelled data for in-context importance.



Optimizing LSA means selecting the appropriate amount of topics. This is usually done by plotting the explained variance ratio and applying the elbow method. LSA doesn't show a clear elbow on our corpus even with a very large number of topics.



We test token retrieval performance at the top-10, 20, 50, and 100 words level using regular term frequency as a baseline. LDA with 10 topics yields the highest recall at every level. tf-idf is also competent.

Integration

Combining Language Importance Scores: We want to incentivize longer sentences in the ranking as very short sentences usually do not exhibit relevant grammatical features. The following formula accumulates importance if a sentence contains multiple lower-frequency words.

$$I_s = \Big(\sum_{w_i \in s} I_{w_i}^4\Big)^{\frac{1}{4}}$$

Combining Corpus Importance Scores: Our implementation of tf-idf allows for zeros and the weights returned by LDA can also be very low. This makes multiplying the scores or adding the log scores intractable. Instead, we take advantage of the normalization and simply take the average.

$$C_s = \frac{\sum_{w_i \in s} C_{w_i}}{|s|}$$

Sentence Ranking with SRS: The final sentence score is a weighted sum of the corpus and language importance scores, which will inform the selection of new cards to learn. It is then used to calculate subsequent intervals for flashcards along with modulators based on a user's answering history.

$$\text{Sentence Score (New Cards)} = \alpha \times \text{Average TF-IDF} + \beta \times \text{DS Score}$$

$$\text{New Interval} = \text{Base Interval} \times (1 + \gamma \times \text{Token Familiarity} + \delta \times \text{Sentence Score})$$

Future Work / References

- Quality Heuristics: Additional sentence ranking heuristics can be implemented based on user feedback.
- Multi-lingual Support: The only language-dependent part in the current implementation is the tokenizer and lemmatizer. We can use a multi-lingual pre-trained model to support parallel corpus between any two languages.
- Improved Topic-Modelling with POS Tagging: Removing words based on POS tags before applying LSA or LDA may provide better retrieval performance.
- Transformer-Based Approach: A possible approach to extract corpus importance score is by varying the attention in transformer-based models. This needs more investigation.

[1] N. J. Cepeda, H. Pashler, E. Vul, J. T. Wixted, and D. Rohrer, "Distributed practice in verbal recall tasks: A review and quantitative synthesis.," Psychological Bulletin, vol. 132, no. 3. American Psychological Association (APA), pp. 354–380, 2006. doi: 10.1037/0033-2909.132.3.354.

[2] A. Tolmachev, S. Kurohashi, and D. Kawahara, "Automatic Japanese Example Extraction for Flashcard-based Foreign Language Learning," Journal of Information Processing, vol. 30, no. 0. Information Processing Society of Japan, pp. 315–330, 2022. doi: 10.2197/ipsjjip.30.315.

[3] P. Lison and J. Tiedemann, 'OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles', in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 923–929.

[4] M. Honnibal and I. Montani, 'spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing'. 2017.

[5] B. Richards, M. H. Daller, D. D. Malvern, P. Meara, J. Milton, and J. Treffers-Daller, Eds., Vocabulary Studies in First and Second Language Acquisition. Palgrave Macmillan UK, 2009. doi: 10.1057/9780230242258.

[6] R. Řehůřek and P. Sojka, 'Software Framework for Topic Modelling with Large Corpora', in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010, pp. 45–50.