

second year project bachelor artificial intelligence



TEAM PERCEPTUM

Recommending document links in the Starfish knowledge graph

Robbert van Ginkel, Jorn Peters & Lotte Weerts

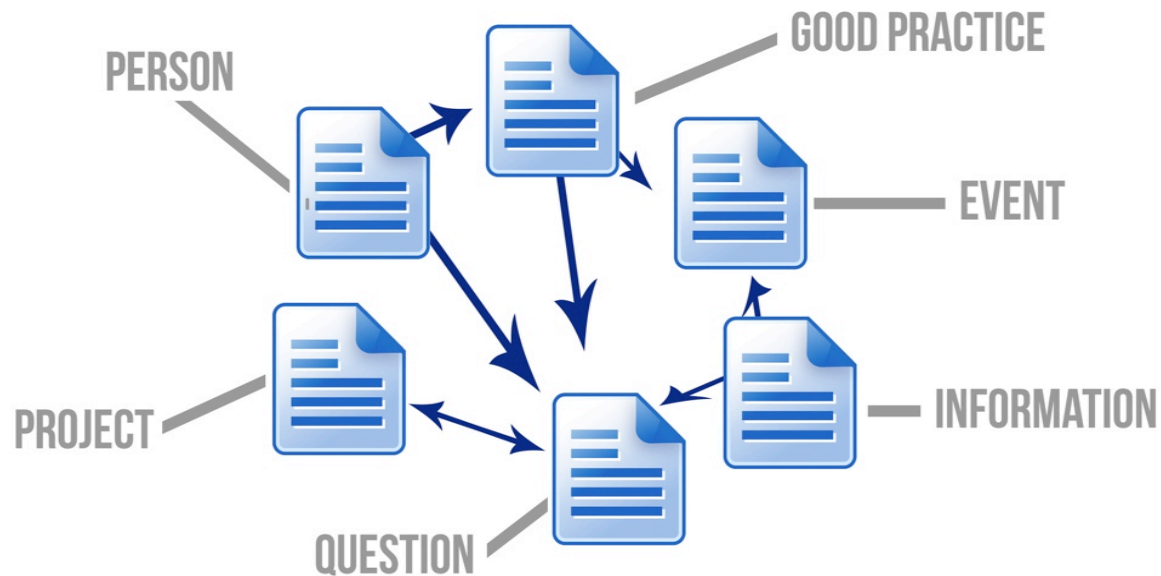
CONTENTS

- Product vision
- Product pipeline
- Evaluation
- Demonstration

PRODUCT VISION

What is Starfish?

Platform for sharing knowledge on education innovation



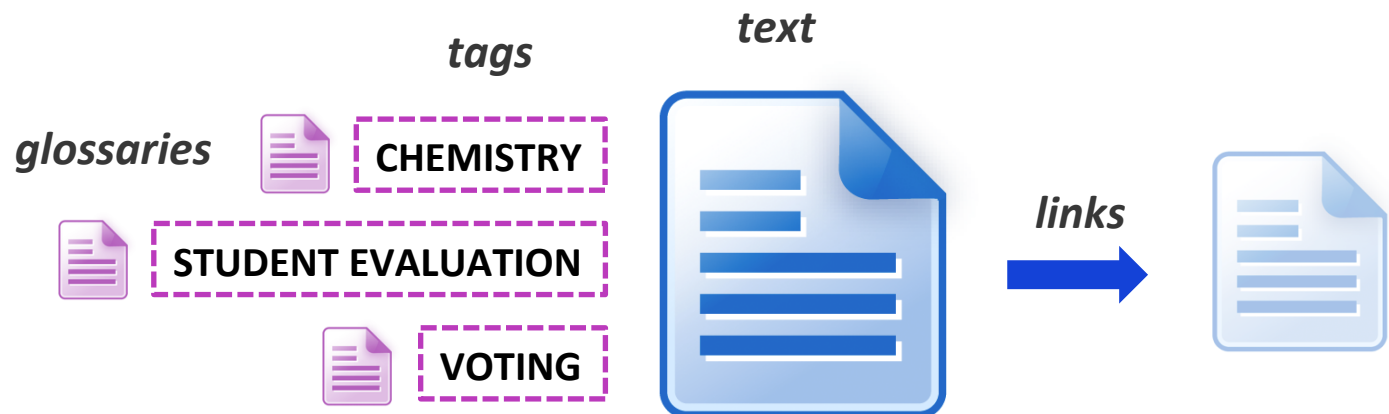
PRODUCT VISION

Product pitch

- **For** Startfish users
- **who** search for and edit knowledge in Starfish
- **the** document linker is a core system addition to StarFish
- **that** finds related documents
- **Unlike** moderated or individual/centralized linking our product uses algorithms and data to automatically suggest document links.

Relevant document properties

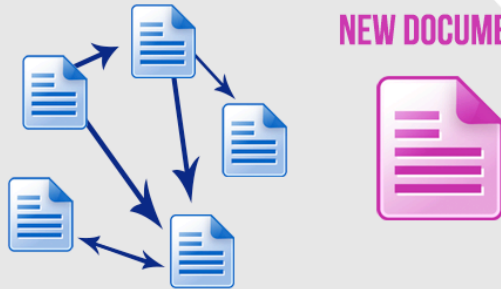
- Textual content of documents
- Tags and their glossaries
- Links to other docs



INPUT

NETWORK

NEW DOCUMENT



OUTPUT

NEW NETWORK



NETWORK DESCRIPTORS

VECTORIZER



NEW DOCUMENT DESCRIPTOR



NEAREST NEIGHBOUR

RANKING

- 1
- 2
- 3
- 4
- 5

THRESHOLD

PROPOSED

- 1
- 2
- 3
- 4
- 5

USER SELECTION

VECTORIZER


$$\begin{bmatrix} 0.003 \\ 0.000 \\ : \\ 0.901 \\ 0.100 \end{bmatrix}$$

➤ **TEXT BASED:** bag of words and TF-IDF

1. Textvectorizer
2. Weighted textvectorizer

➤ **TAG BASED:** occurrences and co-occurrences of tags

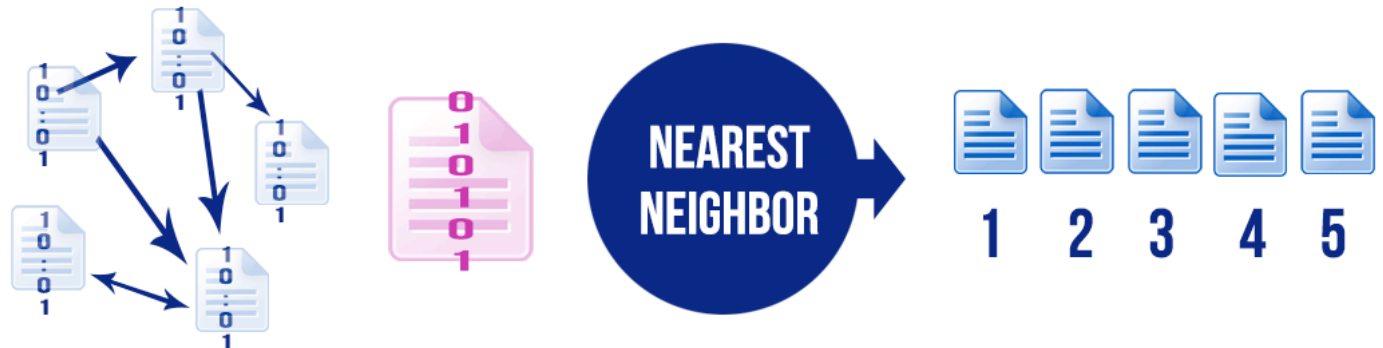
1. Simple tag vectorizer
2. Tag smoothing vectorizer

➤ **HYBRID:** TF-IDF of glossaries of tags

1. Glossaries of tags
2. Weighted glossaries of tags

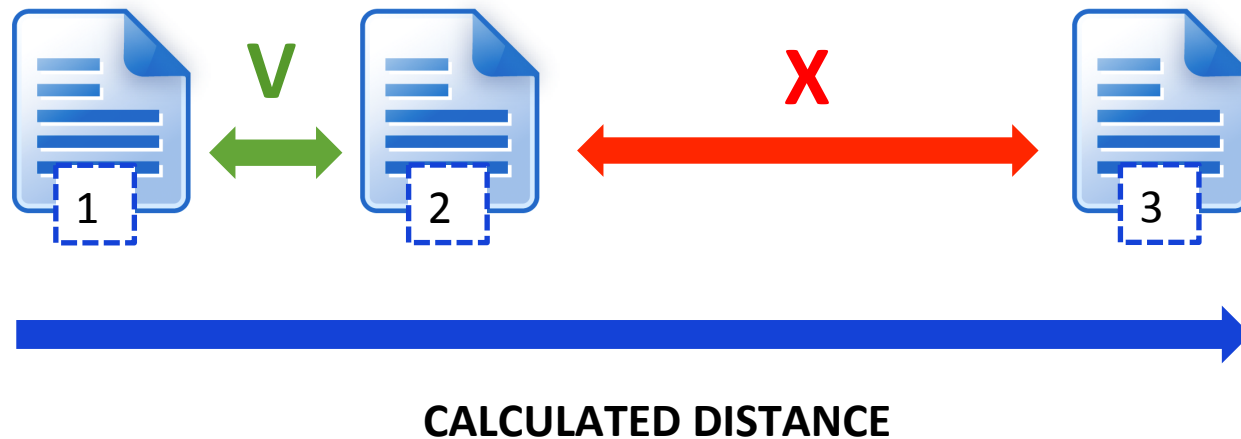
K-NEAREST NEIGHBOR

- Calculate **distance** between descriptor of new document and descriptor of the knowledge base
 1. Cosine
 2. Correlation
- **Rank** documents based on their distances



THRESHOLD

Cut off the number of returned documents based on the **difference between distances** of two consecutive ranks



EVALUATION METRICS

➤ **WITHOUT THRESHOLD:** system returns same amount of links as the document is known to have
Accuracy: correct docs / relevant docs

➤ **WITH THRESHOLD:**

Precision: correct docs / returned docs
User friendliness

Recall: correct docs / relevant docs
Knowledge base coverage

VECTORIZER PERFORMANCE



TABLE

VECTORIZER PERFORMANCE

➤ TEXT BASED

- + xx % accuracy on Questions
- Relatively slow
- Only applicable to textual content
- Bad at handling language differences

➤ TAG BASED:

- + xx % accuracy overall document types
- xx % accuracy on Questions
- Bad performance on no or badly labeled tags

EVALUATION THRESHOLD



TABLE

EVALUATION THRESHOLD

- **COMPARED WITH ACCURACY ON FIXED LINKS:**
 - + Higher precision/recall if tag-based/hybrid
 - Lower recall and precision of text-based

DEMONSTRATION

CONCLUSIONS

- Use *weighted text vectorizer* for Questions
- Use *simple tag vectorizer* for the rest
- Overall performance of entire pipeline:
 - **Precision:** xx% of the recommendations make sense
 - **Recall:** xx% of the relevant documents in the knowledge base are shown

CONCLUSIONS

- Use *weighted text vectorizer* for Questions
- Use *simple tag vectorizer* for the rest
- Overall performance of entire pipeline:
 - **Precision:** xx% of the recommendations make sense
 - **Recall:** xx% of the relevant documents in the knowledge base are shown

ACKNOWLEDGEMENTS

We would like to thank

- Starfish expert **Natasa Brouwer**
- Our academic supervisor **Raquel Fernandez**
- Our clients (but also academic supervisors!)
Robrecht Jurriaans and **Robrecht Jurriaans**



QUESTIONS?

Feel free to ask us!