

second year project bachelor artificial intelligence



# TEAM PERCEPTUM

## *Recommending document links in the Starfish knowledge graph*

Robbert van Ginkel, Jorn Peters & Lotte Weerts

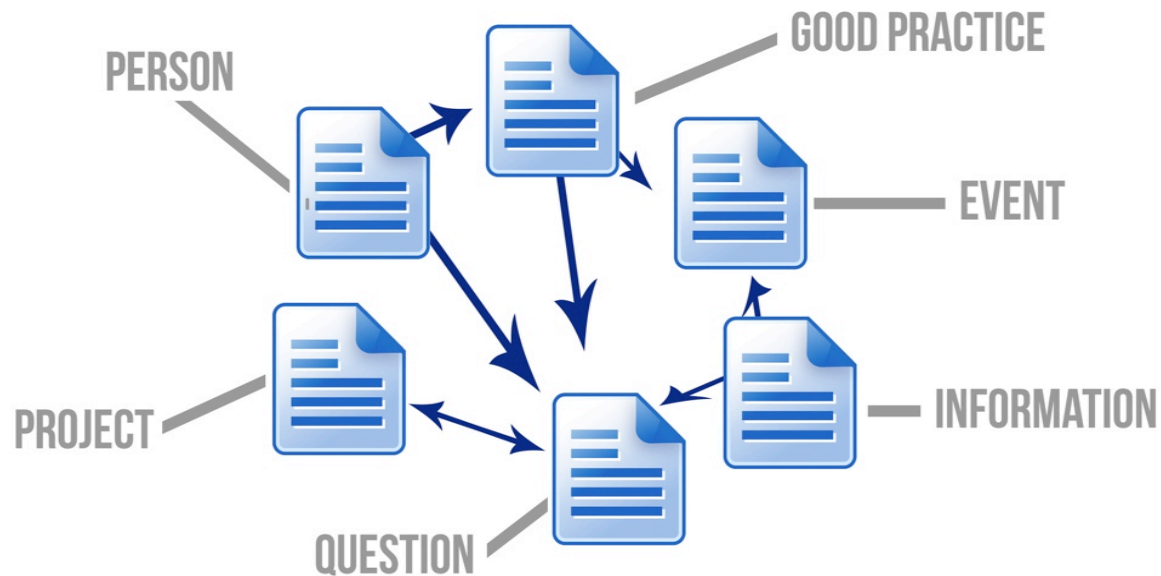
# CONTENTS

- Product vision
- Product pipeline
- Demonstration
- Evaluation
- Conclusion and future work

# PRODUCT VISION

## What is Starfish?

Platform for sharing knowledge on education innovation



# PRODUCT VISION

## Problem

With a set of 200 documents, there are  $2^{(n(n-1)/2)} = 2^{(200*199)/2}$  ways a network can be created of these documents

5045600325728943754639415960323361600171717389128648170285446537790139092022  
4131844468712952801397688747993365836158625307127763731793021606614038184391  
8782166438819034118471727974986376112486958056996757747422199568794123452700  
7218267028650377627355497197592525049358415914726293942898627985869466947873  
8980430166887993069128154835950835018757691057255013110260874429939445352765  
2433382286739063095351150403523845400590029894687157335480176617950005832027  
0062568025185390414082781738607082717107810947125542987701210092607347537510  
0869622598026962991302306673626434584243786355281320519832604041270308253962  
5533250696219082524899445899821285049675946733005361012711646528632594314470  
2466187215344174724752036425156682886050668002000187470606414264577085424670  
5861324179826141955211910198013852756273459089514925741596219256316111451532  
7761664087179834238465797792086513051252298863696094997174378800176909973997

...

# PRODUCT VISION

## Problem

With a set of 200 documents, there are  $2^{(n(n-1)/2)} = 2^{(200*199)/2}$  ways a network can be created of these documents

1140792597464015294976983095494472895933622302169643286731293399566025798222  
9643446091671287381880782103059439224041356391043066777189478984035724387918  
5024375606141333609747594199500471736837358212345193283628673190660152026297  
1823323709392467454084774562469934402259011854550383775670240989737133962322  
8520974685470515471775725205128929319495345324045304939701616981466810498348  
1247298460142120015172985860375751785911442699750496832059717802751704978991  
1329534974491989755960482766136373381395636788077693025438254938790274525059  
8734615460625391992258901631158889652110626047265953663972958603396572930758  
8116345666465743386521886401968708330792381492269069244792180551703743433724  
2525794783317022038171290801260366858253801097560302342031966162233713013755  
8802870126488543467756932095002529646470072800016984930434475473764038329629  
6793840204160947052168071728363833016703797001794142554187375091726258575639

...

# PRODUCT VISION

## Problem

With a set of 200 documents, there are  $2^{(n(n-1)/2)} = 2^{(200*199)/2}$  ways a network can be created of these documents

0780208441706832598565996570394782601338681667703924694035101678176487423014  
9730384096363187661342524339282169862340096839035244942660308392102135985675  
2535451348842002281923789241170546225150321091659509349537054123034949603611  
8053344135612373656880024081361883165025512705396020737495977497506074241370  
4158416749992537522382242326301306866302739979395443292375908934574364490485  
9087420495780395321536687521146231569781390868304660636795034289241589081262  
9574434906254114704295972471412303243282584386318091019559153545707248098969  
6967545022280127135027777604776477790839922208642814953750402954182053977903  
5359951287838870599954969049378833585537427004147826084118548792191034575128  
2207184059589778670185668623196952762298073224336097135238228126937279304379  
1125561161264350238848359428490128197273256366822287909695937801277649059905  
6115271818358337772167604695500813029106234966553281497519213407229933857540

...

# PRODUCT VISION

## Problem

With a set of 200 documents, there are  $2^{(n(n-1)/2)} = 2^{(200*199)/2}$  ways a network can be created of these documents

3230038769840781184454524747638881227563707278138261917498377619244423582082  
4774985002306755907842646253035450988675511553311039082990217919046894581348  
3581073724248265548818214409191591557882710399288043206816393345060354498360  
6320714844930184655509021554230329773972369261139297872135036577653148196200  
9108840518332687536121194234807478596435029262073383582117646070626682341467  
2479306658763127752319142492035107030826495333933315027424605364288301141275  
8276639065671168810639472762789872319982092821807253584095799112450096211555  
8117060795881832099007064489924753846648178695164000823527116033763652508927  
8334014133866301196608594779157588745038697750058077049253527187367476153890  
7957592692338873200981438032575284132241766498507615344640955103466518325731  
7962315817338581050676667855409763833079240562945202880215897410960959171497  
2544655840163092845489531609241042878137835728275791923745816998572435391873

...

# PRODUCT VISION

## Problem

With a set of 200 documents, there are  $2^{(n(n-1)/2)} = 2^{(200*199)/2}$  ways a network can be created of these documents

4595320465324426949175556678205791255375018617267121794252557026943573585031  
8452705408299142586139212884744497347190160137840419405247039518396274114645  
2303996517847513395369655782580523508813546833814650064457048706912704603464  
9757504890757064085820563631196750640718045809549143688428793431089484969523  
1749241594632170370323090747712206506475569253301878996365153005838577356222  
0560563090153780856286764728881037486300365506530567718047452433561939929516  
3746540550242704334913580642576490058983782054522498658650750923394970585810  
5241239788158201196922887758490437884852932384943096357852846868004827314575  
5569627169424695260812022089918983177038899603863630222342822859809610153359  
0272587847048035742540101250751945672774500879564871885158585713790402016379  
1616285476664724474202618212705667501028094222812833917300899405839110461961  
5918532519310724925080940979420384824791211727725922911055215410441720330007

...



# PRODUCT VISION

## Problem

With a set of 200 documents, there are  $2^{(n(n-1)/2)} = 2^{(200*199)/2}$  ways a network can be created of these documents

2816826377953293734622839346686527363977978421772469603279239571026523518964  
7100569957347321991434552856120268551577844079284194220801575977364202380352  
2642892349322328122509997367290023985615538788322920369542996134468029282460  
5921998426036928632274408900497201850210335685075124559455805238390219799915  
7806042293053500470907202978672940884840287066626585954456983413233103495280  
4602986746268651256453906106915792705500520667940910917258510220425892034109  
9246612675684773176746196202085856935671553979512532931263612306927801751704  
9691460429829011293048981909027135397245002114181195048326550888139815883572  
9275508287833977938697072866989761789024794395522593011182756458764834579487  
7797582577031372273032590118758826814689728989735427869363448961435714259288  
3560351408529763186276228169485308969005836682216180657362164278092684448923  
4549137636304508299928575502337178321772555196881876674292971429886378857453

...

# PRODUCT VISION

## Problem

With a set of 200 documents, there are  $2^{n(n-1)/2} = 2^{(200*199)/2}$  ways a network can be created of these documents

2447923933083569108416434916530837313030654222351593453486277619448383035960  
8113811349952236170973177022980302338170846856335599860080992654070123610237  
3276209666653184120118893382218875940655293277957600583339656379698679560434  
0884677170302507419405754082096159628786799476560747413313082774756353231235  
2079647668069335355161195795417508584473346347967227349546355337402005817188  
72004005783138992282399118749398710969583868.

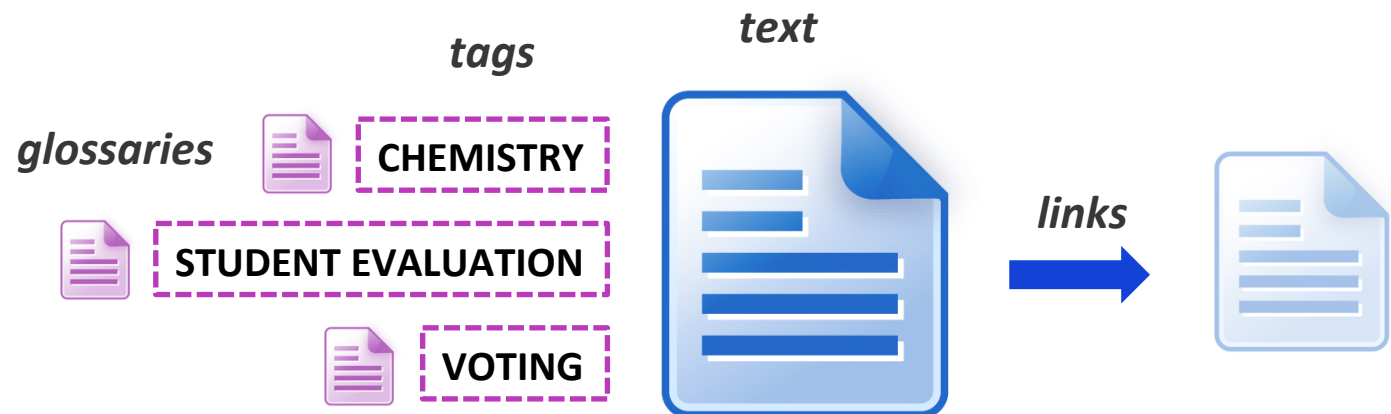
# PRODUCT VISION

## Product pitch

- **For** Starfish 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

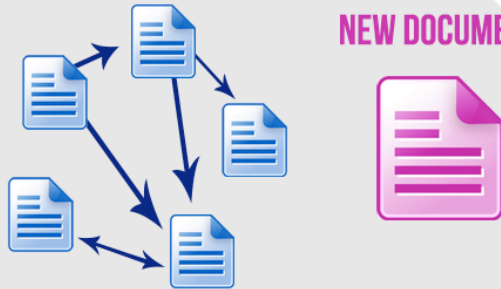
- Textual content of documents
- Tags and their glossaries
- Links to other docs



# 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



[ 0.003  
0.000  
:  
0.901  
0.100 ]

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

1. Textvectorizer
2. Weighted textvectorizer

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

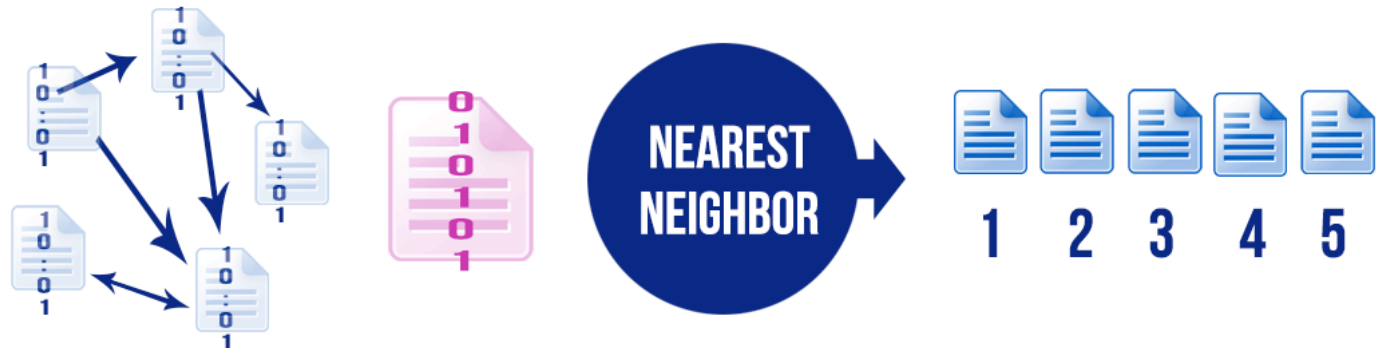
1. Simple tag vectorizer
2. Tag smoothing vectorizer

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

1. Glossaries of tags
2. Weighted glossaries of tags

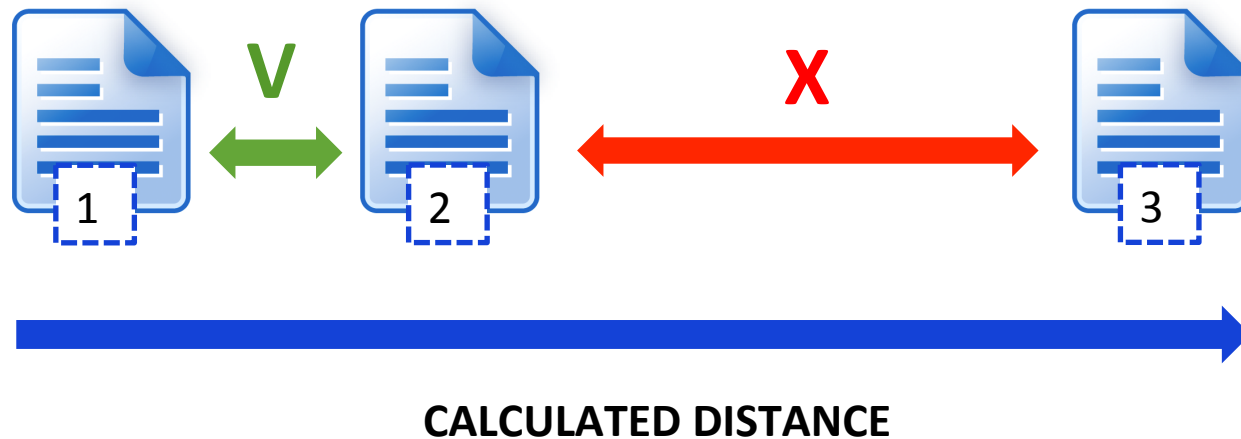
# K-NEAREST NEIGHBOR

- 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





# DEMONSTRATION OF OUTPUT

## Performance report

%%%

Performance report

%%%

Average recall: 0.4972377311162357891329853946

Average precision: 0.5093457943925233644859813083

Average recall per type

Information: 0.5497453526865291571173924112

Question: 0.3964912280701754385964912281

Good Practice: 0.3214285714285714285714285715

Project: 0.307291666666666666666666666666

Person: 0.5820512820512820512820512821

Event: 0.1785714285714285714285714286

Average precision per type

Information: 0.6519607843137254901960784312

Question: 0.50

Good Practice: 0.375

Project: 0.625

Person: 0.3931623931623931623931623931

Event: 0.333333333333333333333333333333

# DEMONSTRATION OF OUTPUT

## HTML webpage

### 27 - TPACK - E-Learning Cookbook (Information)

By Natasa Brouwer

Content

Links

#### ♦ Natasa Brouwer (Person)

ActiveLearning

Natuurwetenschappen

AfstandOnderwijsEnZelfstandigLeren

Docentprofessionalisering

Blackboard

ECTN

Chemistry

Stemkastjes

ToetsenEnToetsgestuurdLeren

Content

#### ♦ Andr Heck (Person)

MapleTA

Natuurwetenschappen

AfstandOnderwijsEnZelfstandigLeren

LearningAnalytics

DigitalAssessmentTools

ToetsenEnToetsgestuurdLeren

Content

#### ♦ Erwin van Vliet (Person)

Psychobiology

AfstandOnderwijsEnZelfstandigLeren

Stemkastjes

ActiveLearning

Think-pair-share

FlippedClassroom

#### ♦ Wat is het verschil tussen Learning Analytics en TTL (Question)

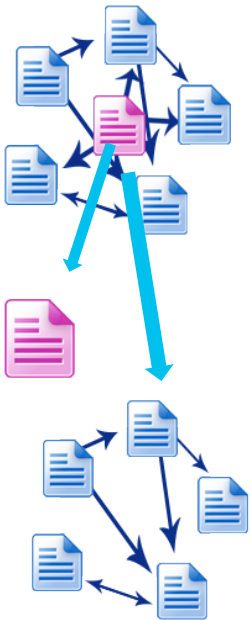
ToetsenEnToetsgestuurdLeren

LearningAnalytics

By Natasa Brouwer

Content

# EVALUATION METRICS



➤ TAKE ONE OUT PRINCIPLE

➤ WITH THRESHOLD WE CAN MEASURE:

**Precision:**  $\frac{|correct\ proposed\ docs|}{|proposed\ documents|}$  *(user friendliness)*

**Recall:**  $\frac{|correct\ proposed\ docs|}{|relevant\ documents|}$  *(corpus coverage)*

**F1-Measure:**  $\frac{2 * precision * recall}{precision + recall}$  *(trade-off precision and recall)*

# PERFORMANCE

Recall

Precision

Vectorizer	Info.	Question	Good Pr.	Project	Person	Event	Average	F1
Text	12.1	39.6	19.6	20.7	5.2	21.4	15.66	18.97
	26.5	50.0	41.7	24.8	24.8	7.3	24.05	
Weighted text	14.8	29.8	19.6	24.4	5.2	21.4	15.11	18.23
	26.8	41.7	33.3	26.3	8.6	27.8	23.00	
Simple tag	55.0	20.6	32.1	30.7	58.2	17.9	46.34	45.71
	64.2	17.1	37.5	62.5	39.3	33.3	45.09	
Tag smoothing	55.6	20.6	42.9	34.9	66.3	33.7	49.56	43.20
	46.1	18.6	43.8	56.3	36.3	44.4	38.29	
Glossaries of tags	36.5	23.3	21.4	36.4	50.2	44.1	38.80	28.08
	25.0	14.5	37.5	33.2	17.3	46.7	22.00	
Weighted tags	36.5	23.3	21.4	36.4	50.2	44.1	38.80	28.08
	25.0	14.5	37.5	33.2	17.33	46.7	22.00	

# PERFORMANCE

Recall

Precision

Vectorizer	Info.	Question	Good Pr.	Project	Person	Event	Average	F1
Text	12.1	39.6	19.6	20.7	5.2	21.4	15.66	18.97
	26.5	50.0	41.7	24.8	24.8	7.3	24.05	
Weighted text	14.8	29.8	19.6	24.4	5.2	21.4	15.11	18.23
	26.8	41.7	33.3	26.3	8.6	27.8	23.00	
Simple tag	55.0	20.6	32.1	30.7	58.2	17.9	46.34	45.71
	64.2	17.1	37.5	62.5	39.3	33.3	45.09	
Tag smoothing	55.6	20.6	42.9	34.9	66.3	33.7	49.56	43.20
	46.1	18.6	43.8	56.3	36.3	44.4	38.29	
Glossaries of tags	36.5	23.3	21.4	36.4	50.2	44.1	38.80	28.08
	25.0	14.5	37.5	33.2	17.3	46.7	22.00	
Weighted tags	36.5	23.3	21.4	36.4	50.2	44.1	38.80	28.08
	25.0	14.5	37.5	33.2	17.33	46.7	22.00	

# PERFORMANCE

Recall

Precision

Vectorizer	Info.	Question	Good Pr.	Project	Person	Event	Average	F1
Text	12.1	39.6	19.6	20.7	5.2	21.4	15.66	18.97
	26.5	50.0	41.7	24.8	24.8	7.3	24.05	
Weighted text	14.8	29.8	19.6	24.4	5.2	21.4	15.11	18.23
	26.8	41.7	33.3	26.3	8.6	27.8	23.00	
Simple tag	55.0	20.6	32.1	30.7	58.2	17.9	46.34	45.71
	64.2	17.1	37.5	62.5	39.3	33.3	45.09	
Tag smoothing	55.6	20.6	42.9	34.9	66.3	33.7	49.56	43.20
	46.1	18.6	43.8	56.3	36.3	44.4	38.29	
Glossaries of tags	36.5	23.3	21.4	36.4	50.2	44.1	38.80	28.08
	25.0	14.5	37.5	33.2	17.3	46.7	22.00	
Weighted tags	36.5	23.3	21.4	36.4	50.2	44.1	38.80	28.08
	25.0	14.5	37.5	33.2	17.33	46.7	22.00	

# VECTORIZER PERFORMANCE

## ➤ TEXT BASED

- + 39.6% recall and 50% precision on Questions
- Relatively slow
- Only applicable to textual content
- Bad at handling language differences

## ➤ TAG BASED:

- + 45.71% F-1 overall document types
- 20.6% recall 17.1% precision on Questions
- Bad performance on no or badly labeled tags

# VECTORIZER PERFORMANCE

## HYBRID TEXTVECTORIZER & SIMPLE TAG VECTORIZER

Vectorizer	Recall		Precision				Average	F1
	Info.	Question	Good Pr.	Project	Person	Event		
Hybrid	55.0	39.6	32.1	30.7	58.2	17.9	49.72	50.32
	64.2	50.0	37.5	62.5	39.3	33.3	50.93	

↑  
Textvectorizer



# CONCLUSIONS

- Use *text vectorizer* for Questions
- Use *simple tag vectorizer* for the rest
- Overall performance of entire pipeline:
  - **Precision:** 50.93% of the recommendations make sense
  - **Recall:** 49.72% of the relevant documents in the knowledge base are shown

# FUTURE WORK

- Links in Starfish are directed, but now only outgoing links are proposed. Incoming links should also be proposed.
- Calculate link-probabilities if a larger data set is available
- Use LDA (*Latent Dirichlet Allocation*) to generate topics if a document has no tags

# ACKNOWLEDGEMENTS

We would like to thank

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

# QUESTIONS?

Feel free to ask us!