second year project bachelor artificial intelligence

TEAM PERCEPTUM

Recommending document links in the Starfish knowledge graph

Robbert van Ginkel, Jorn Peters & Lotte Weerts

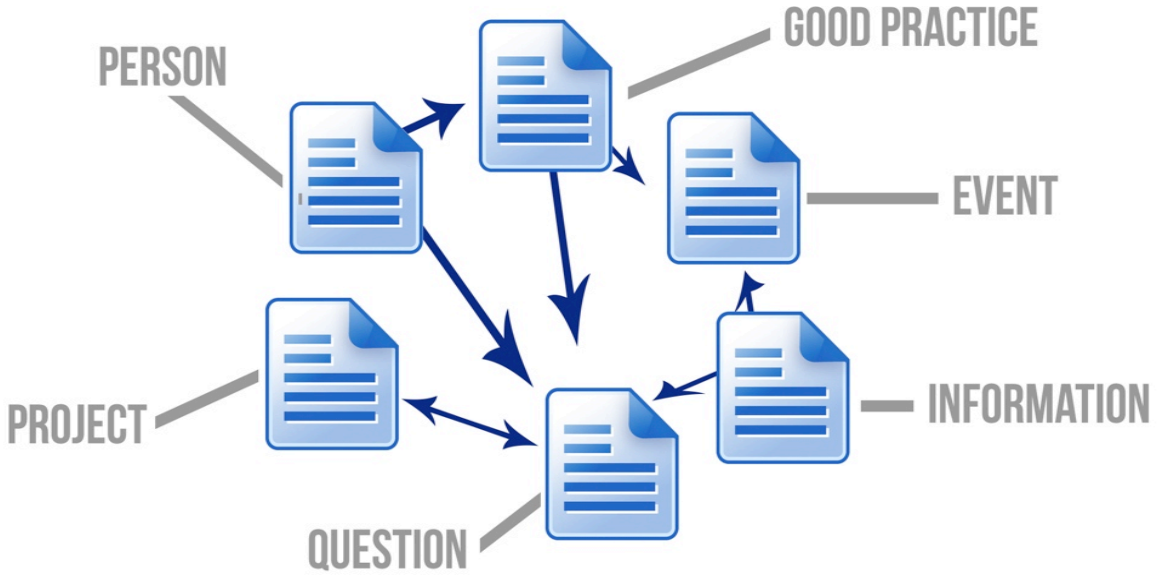
CONTENTS

- Product vision
- Product pipeline
- Demonstration
- Evaluation

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

50456003257289437546394159603233616001717173891286481702854465377901390920224131844468712952801397688747993365836158625307127763731793021606614038184391878216643881903411847172797498637611248695805699675774742219956879412345270072182670286503776273554971975925250493584159147262939428986279858694669478738980430166887993069128154835950835018757691057255013110260874429939445352765243338228673906309535115040352384540059002989468715733548017661795000583202700625680251853904140827817386070827171078109471255429877012100926073475375100869622598026962991302306673626434584243786355281320519832604041270308253962553325069621908252489944589982128504967594673300536101271164652863259431447024661872153441747247520364251566828860506680020001874706064142645770854246705861324179826141955211910198013852756273459089514925741596219256316111451532776166408717983423846579779208651305125229886369609499717437880017690997399…

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

07802084417068325985659965703947826013386816677039246940351016781764874230149730384096363187661342524339282169862340096839035244942660308392102135985675253545134884200228192378924117054622515032109165950934953705412303494960361180533441356123736568800240813618831650255127053960207374959774975060742413704158416749992537522382242326301306866302739979395443292375908934574364490485908742049578039532153668752114623156978139086830466063679503428924158908126295744349062541147042959724714123032432825843863180910195591535457072480989696967545022280127135027777604776477790839922208642814953750402954182053977903535995128783887059995496904937883358553742700414782608411854879219103457512822071840595897786701856686231969527622980732243360971352382281269372793043791125561161264350238848359428490128197273256366822287909695937801277649059905611527181835833772167604695500813029106234966553281497519213407229933857540…

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

32300387698407811844545247476388812275637072781382619174983776192444235820824774985002306755907842646253035450988675511553311039082990217919046894581348358107372424826554881821440919159155788271039928804320681639334506035449836063207148449301846555090215542303297739723692611392978721350365776531481962009108840518332687536121194234807478596435029262073383582117646070626682341467247930665876312775231914249203510703082649533393331502742460536428830114127582766390656711688106394727627898723199820928218072535840957991124500962115558117060795881832099007064489924753846648178695164000823527116033763652508927833401413386630119660859477915758874503869775005807704925352718736747615389079575926923388732009814380325752841322417664985076153446409551034665183257317962315817338581050676667855409763833079240562945202880215897410960959171497254465584016309284548953160924104287813783572827579192374581699857243539187325…

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

459532046532442694917555667820579125537501861726712179425255702694357358503184527054082991425861392128847444973471901601378404194052470395183962741146452303996517847513395369655782580523508813546833814650064457048706912704603464975750489075706408582056363119675064071804580954914368842879343108948496952317492415946321703703230907477122065064755692533018789963651530058385773562220560563090153780856286764728881037486300365506530567718047452433561939929516374654055024270433491358064257649005898378205452249865865075092339497058581052412397881582011969228877584904378848529323849430963578528468680048273145755569627169424695260812022089918983177038899603863630222342822859809610153359027258784704803574254010125075194567277450087956487188515858571379040201637916162854766647244742026182127056675010280942228128339173008994058391104619615918532519310724925080940979420384824791211727725922911055215410441720330007…

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
7200400578313899228239911874939871096958386 8.

PRODUCT VISION

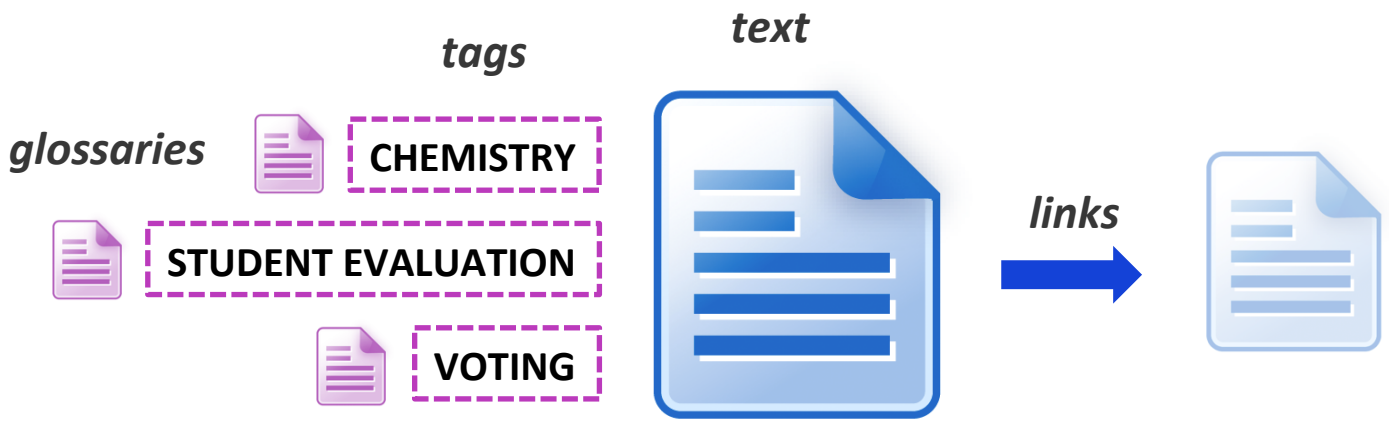
Product pitch

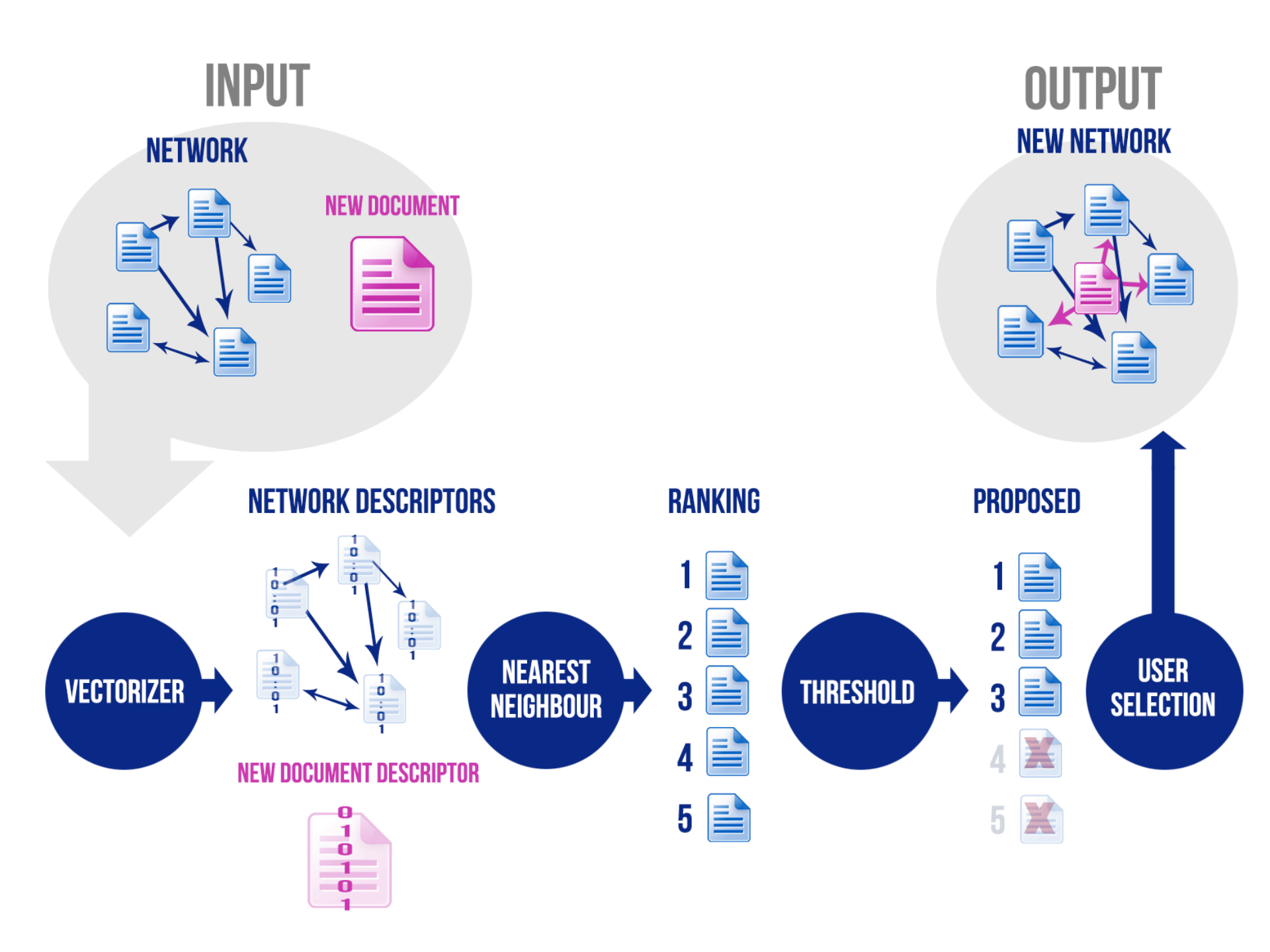
- **For** Starfish users
- **who** search for and edit knowledge in Starfish
- **the** document linker is a core system addition to Starfish
- **that** finds related documents
- **Unlike** moderated or individual/centralized linking our product uses algorithms and data to automatically suggest document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

DOMAIN

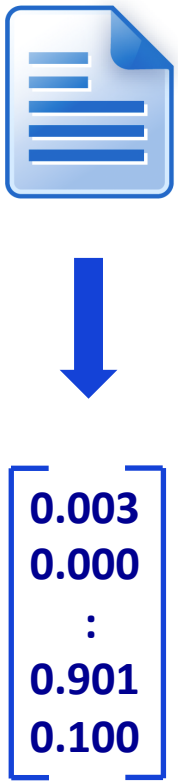
Relevant document properties

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




INPUT
NETWORK
NEW DOCUMENT
OUTPUT
NEW NETWORK
NETWORK DESCRIPTORS
VECTORIZER
NEAREST NEIGHBOUR
RANKING
1
2
3
4
5
THRESHOLD
PROPOSED
1
2
3
4
5
USER SELECTION
NEW DOCUMENT DESCRIPTOR

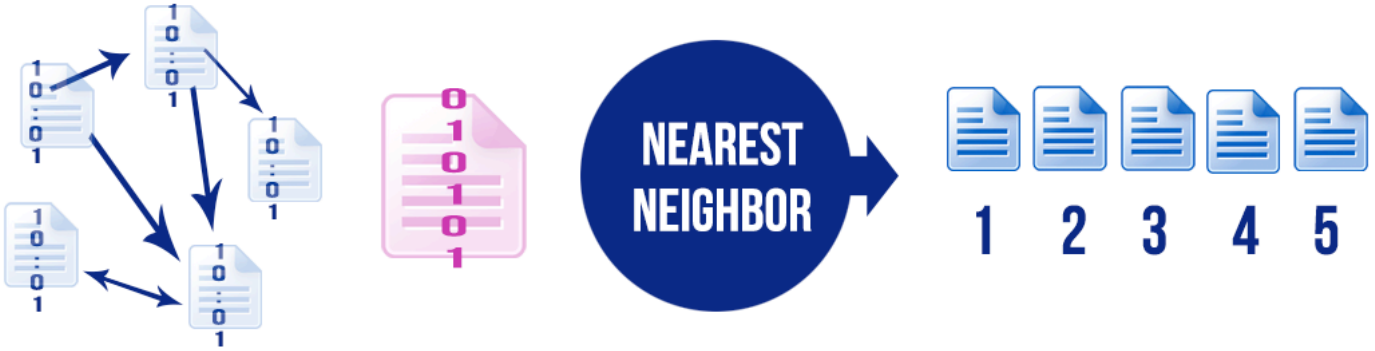
VECTORIZER



- **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

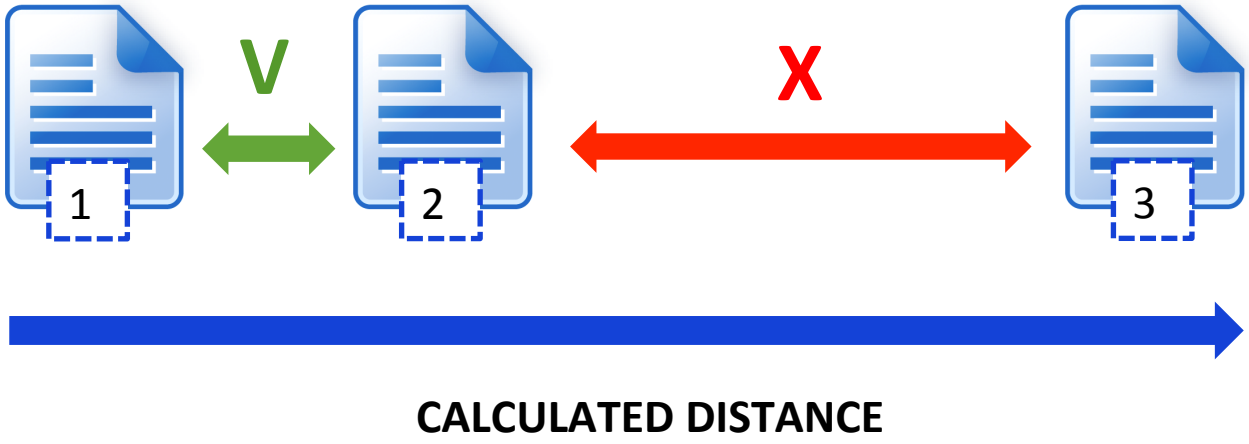
K-NEAREST NEIGHBOR

- 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

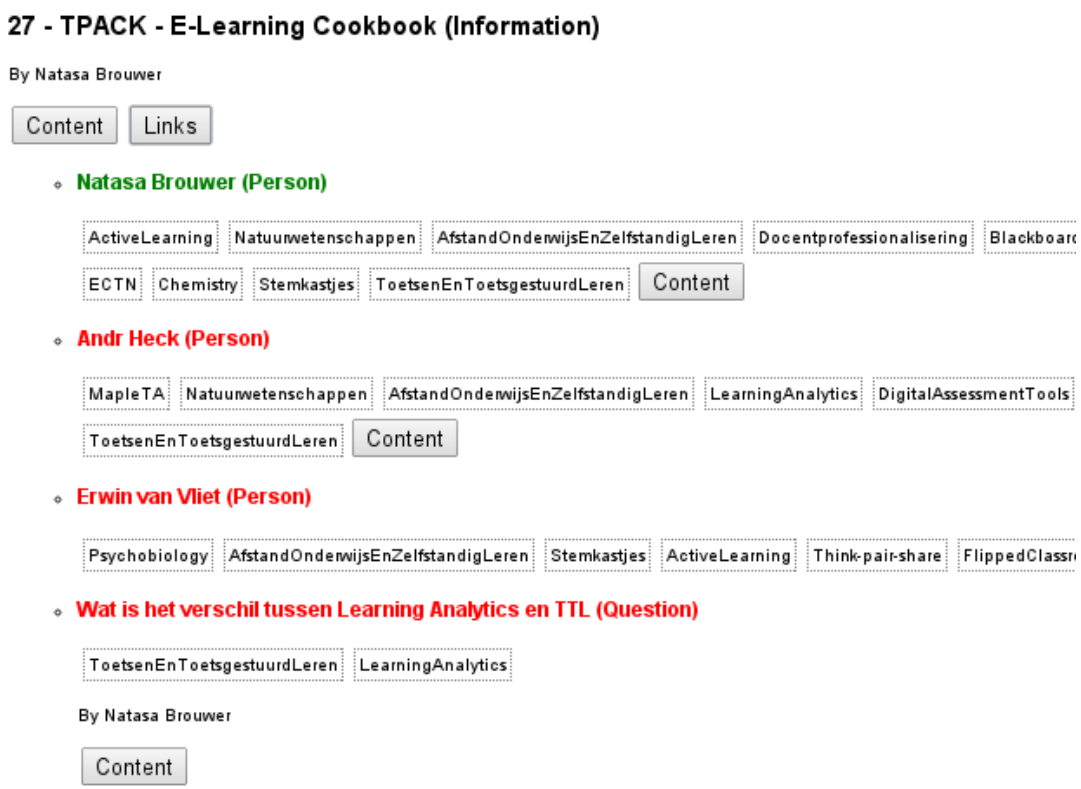
Information:	0.5497453526865291571173924112
Question:	0.3964912280701754385964912281
Good Practice:	0.3214285714285714285714285715
Project:	0.3072916666666666666666666666
Person:	0.5820512820512820512820512821
Event:	0.1785714285714285714285714286

Average precision per type

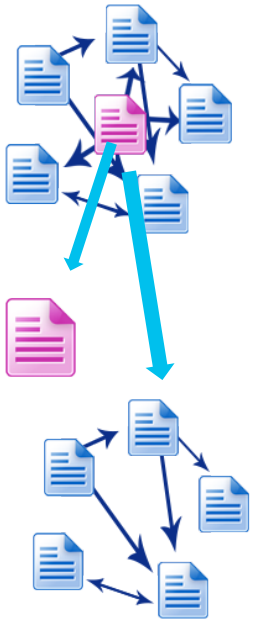
Information:	0.6519607843137254901960784312
Question:	0.50
Good Practice:	0.375
Project:	0.625
Person:	0.3931623931623931623931623931
Event:	0.3333333333333333333333333333

DEMONSTRATION OF OUTPUT

HTML webpage



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

Recall **Precision**

Vectorizer	Info.	Question	Good Pr.	Project	Person	Event	Average	F1
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 Dichliret Allocation*) to generate topics if a document has no tags

ACKNOWLEDGEMENTS

We would like to thank

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

QUESTIONS?

Feel free to ask us!