CONTENTS
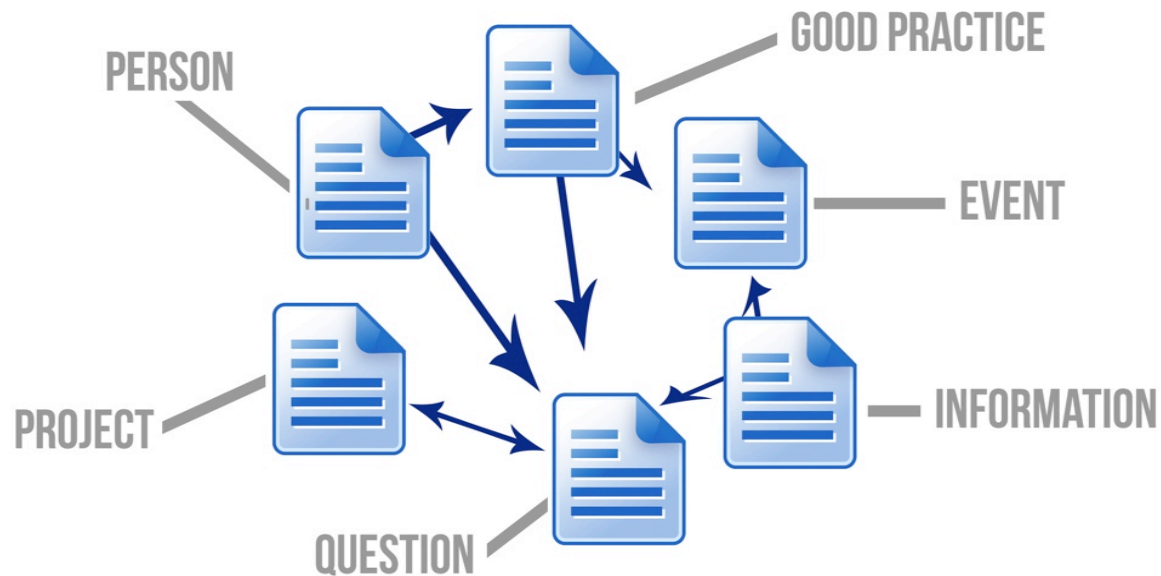
- ↗ Product vision

- ↗ Product pipeline

- ↗ Demonstration

- ↗ Evaluation

What is Starfish?

Platform for sharing knowledge on education innovation

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 7

…

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

11407925974640152949769830954944728959336223021696432867312933995660257982229643446091671287381880782103059439224041356391043066777189478984035724387918502437560614133360974759419950047173683735821234519328362867319066015202629718233237093924674540847745624699344022590118545503837756702409897371339623228520974685470515471775725205128929319495345324045304939701616981466810498348124729846014212001517298586037575178591144269975049683205971780275170497899113295349744919897559604827661363733813956367880776930254382549387902745250598734615460625391992258901631158889652110626047265953663972958603396572930758811634566646574338652188640196870833079238149226906924479218055170374343372425257947833170220381712908012603668582538010975603023420319661622337130137558802870126488543467756932095002529646470072800016984930434475473764038329629679384020416094705216807172836383301670379700179414255418737509172625857563 9 ...

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
1125561161264350238848359428490128197273256366822879096959378012776490599056
1152718183583377721676046955008130291062349665532814975192134072299338857540
...

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

32300387698407811844545247476388812275637072781382619174983776192444235820824774985002306755907842646253035450988675511553311039082990217919046894581348358107372424826554881821440919159155788271039928804320681639334506035449836063207148449301846555090215542303297739723692611392978721350365776531481962009108840518332687536121194234807478596435029262073383582117646070626682341467247930665876312775231914249203510703082649533393331502742460536428830114127582766390656711688106394727627898723199820928218072535840957991124500962115558117060795881832099007064489924753846648178695164000823527116033763652508927833401413386630119660859477915758874503869775005807704925352718736747615389079575926923388732009814380325752841322417664985076153446409551034665183257317962315817338581050676667855409763833079240562945202880215897410960959171497254465584016309284548953160924104287813783572827579192374581699857243539187 3 …

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

459532046532442694917555667820579125537501861726712179425257026943573585031
845270540829914258613921288474449734719016013784041940524703951839627411464523
039965178475133953696557825805235088135468338146500644570487069127046034649
7575048907570640858205636311967506407180458095491436884287934310894849695231
749241594632170370323090747712206506475569253301878996365153005838577356222
056056309015378085628676472888103748630036550653056771804745243356193992951637
465405502427043349135806425764900589837820545224986586507509233949705858105
24123978815820119692288775849043788485293238494309635785284686800482731457555
696271694246952608120220899189831770388996038636302223428228598096101533590
2725878470480357425401012507519456727745008795648718851585857137904020163791
616285476647244742026182127056675010280942228128339173008994058391104619615
918532519310724925080940979420384824791211727725922911055215410441720330007
…

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
7806042293053500470907202978672940884840287066262658595445698341323310349528
0
4602986746268651256453906106915792705500520667940910917258510220425892034109
9246612675684773176746196202085856935671553979512532931263612306927801751704
9691460429829011293048981909027135397245002114181195048326550888139815883572
9275508287833977938697072866989761789024794395522593011182756458764834579487
7797582577031372273032590118758826814689728989735427869363448961435714259288
3560351408529763186276228169485308969005836682216180657362164278092684448923
4549137636304508299928575502337178321772555196881876674292971429886378857453
…

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

244792393308356910841643491653083731303065422235159345348627761944838303596081138113499522361709731770229803023381708468563355998600809926540701236102373276209666653184120118893382188759406552932779576005833396563796986795604340884677170302507419405754082096159628786799476560747413313082774756353231235207964766806933535516119579541750858447334634796722734954635533740200581718872004005783138992282399118749398710969583868.

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**
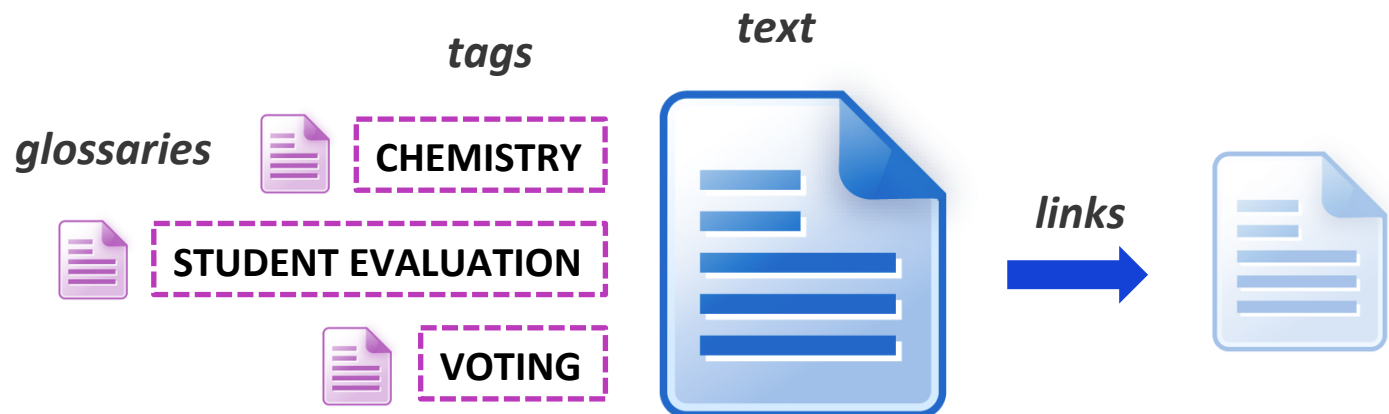
INPUT

NETWORK

NEW DOCUMENT
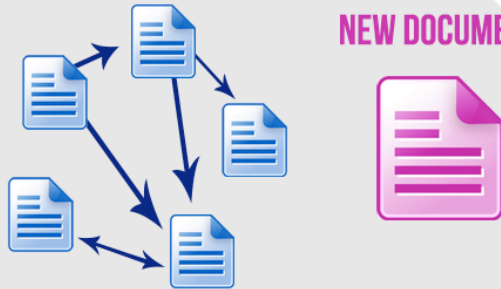
OUTPUT

NEW NETWORK

NETWORK DESCRIPTORS

RANKING

PROPOSED

VECTORIZER

NEW DOCUMENT DESCRIPTOR

NEAREST NEIGHBOUR

1
2
3
4
5

THRESHOLD

1
2
3
4
5

USER SELECTION

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

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

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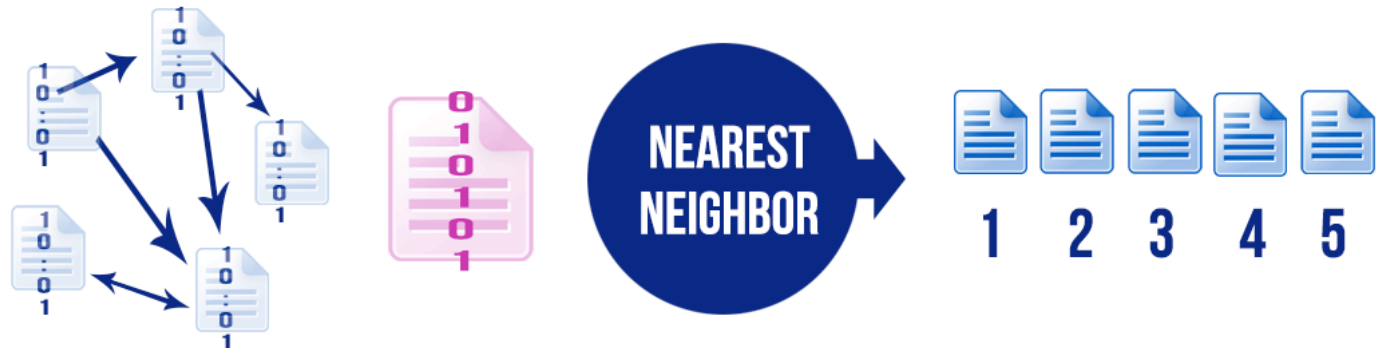
Cut off the number of returned documents based on the **difference between distances** of two consecutive ranks

CALCULATED DISTANCE

DEMONSTRATION OF OUTPUT

Performance report

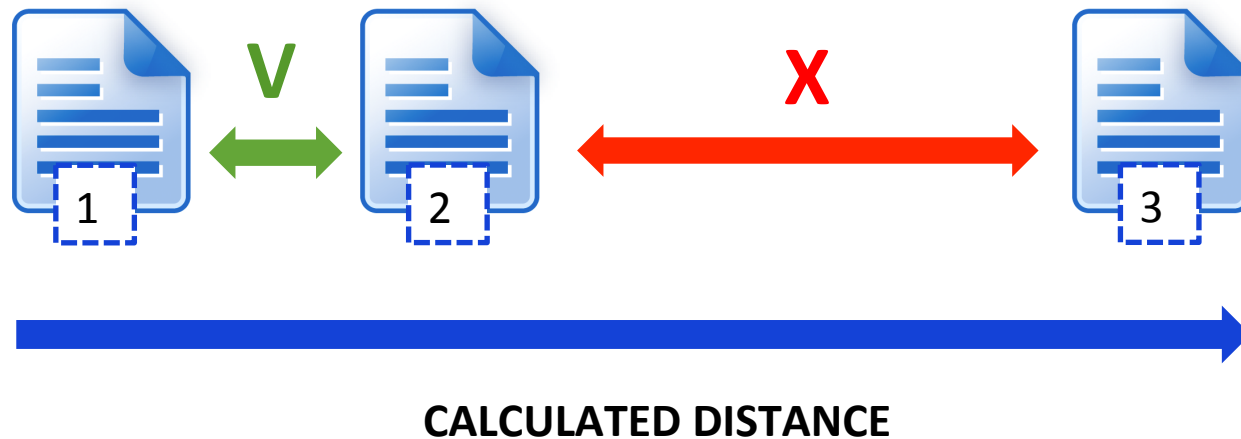%%%
Performance report
%%%

Average recall: 0.49723773111623578913298539461
Average precision: 0.50934579439252336448598130883

Average recall per type
Information: 0.54974535268652915711739241112
Question: 0.39649122807017543859649122281
Good Practice: 0.32142857142857142857142857142715
Project: 0.30729166666666666666666666666
Person: 0.58205128205128205128205128211
Event: 0.17857142857142857142857142857142857144286

Average precision per type
Information: 0.65196078431372549019607843112
Question: 0.50
Good Practice: 0.375
Project: 0.625
Person: 0.39316239316239316239316239316239316239311
Event: 0.33333333333333333333333333333333

DEMONSTRATION OF OUTPUT

HTML webpage

EVALUATION METRICS

↗ **TAKE ONE OUT PRINCIPLE**

↗ **WITH THRESHOLD WE CAN MEASURE:**

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

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

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

PERFORMANCE

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**	

Recall

Precision

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

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!