CONTENTS
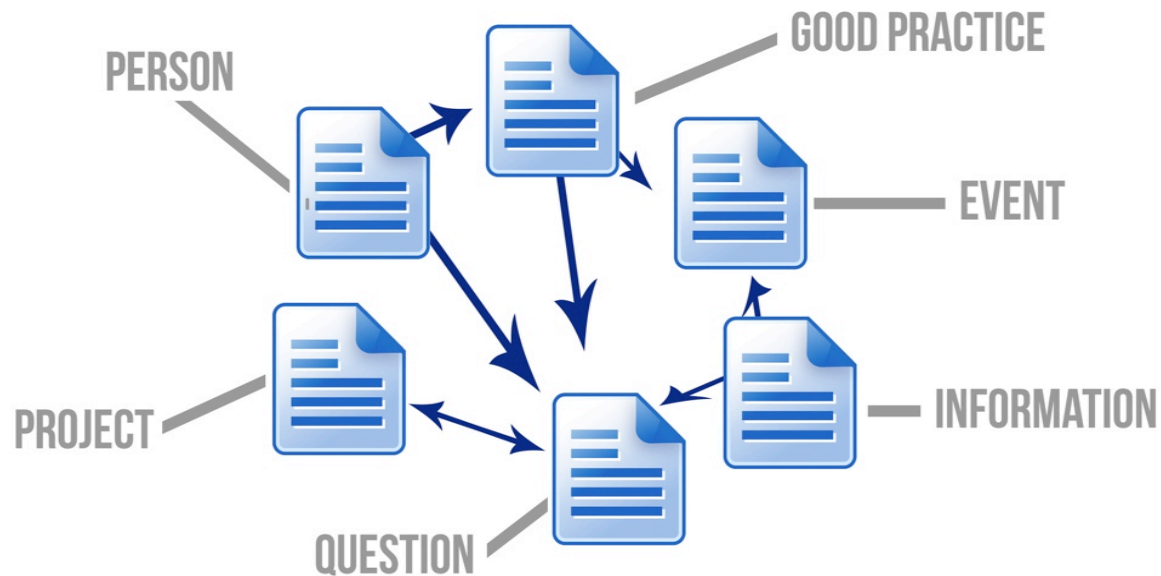
- ↗ Product vision
- ↗ Product pipeline
- ↗ Evaluation
- ↗ Demonstration

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

…

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
…

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

07802084417068325985659965703947826013386816677039246940351016781764874230149730384096363187661342524339282169862340096839035244942660308392102135985675253545134884200228192378924117054622515032109165950934953705412303494960361180533441356123736568800240813618831650255127053960207374959774975060742413704158416749992537522382242326301306866302739979395443292375908934574364490485908742049578039532153668752114623156978139086830466063679503428924158908126295744349062541147042959724714123032432825843863180910195591535457072480989696967545022280127135027777604776477790839922208642814953750402954182053977903535995128783887059995496904937883358553742700414782608411854879219103457512822071840595897786701856686231969527622980732243360971352382281269372793043791125561161264350238848359428490128197273256366822879096959378012776490599056115271818358337772167604695500813029106234966553281497519213407229933857540
…

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
…

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
845270540829914258613921288474449734719016013784041940524703951839627411464
523039965178475133953696557825805235088135468338146500644570487069127046034
649757504890757064085820563631196750640718045809549143688428793431089484969523
174924159463217037032309074771220650647556925330187899636515300583857735622
205605630901537808562867647288810374863003655065305677180474524335619399295
163746540550242704334913580642576490058983782054522498658650750923394970585810
524123978815820119692288775849043788485293238494309635785284686800482731457
556962716942469526081202208991898317703889960386363022234282285980961015335
902725878470480357425401012507519456727745008795648718851585857137904020163799
161628547666472447420261821270566750102809422812833917300899405839110461961
591853251931072492508094097942038482479121172772592291105521541044172033000
…

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

281682637795329373462283934668652736397797842177246960327923957102652351896471005699573473219914345528561202685515778440792841942208015759773642023803522642892349322328122509997367290023985615538788322920369542996134468029282460592199842603692863227440890049720185021033568507512455945580523839021979991578060422930535004709072029786729408848402870666265859544569834132331034952804602986746268651256453906106915792705500520667940910917258510220425892034109924661267568477317674619620208585693567155397951253293126361230692780175170496914604298290112930489819090271353972450021141811950483265508881398158835729275508287833977938697072866989761789024794395522593011182756458764834579487779758257703137227303259011875882681468972898973542786936344896143571425928835603514085297631862762281694853089690058366822161806573621642780926844489234549137636304508299928575502337178321772555196881876674292971429886378857453

…

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

24479239330835691084164349165308373130306542223515934534862776194483830359608113811349952236170973177022980302338170846856335599860080992654070123610237327620966665318412011889338221887594065529327795760058333965637969867956043408846771703025074194057540820961596287867994765607474133130827747563532312352079647668069335355161195795417508584473346347967227349546355337402005817188720040057831389922823991187493987109695838688.

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.
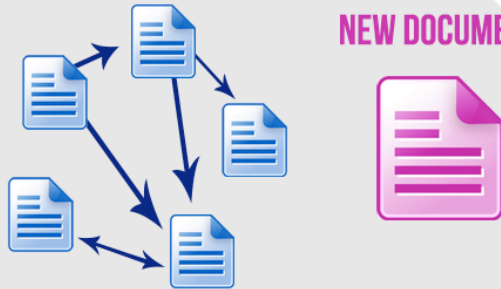
Relevant document properties

↗ **Textual content of documents**

↗ **Tags and their glossaries**

↗ **Links to other docs**
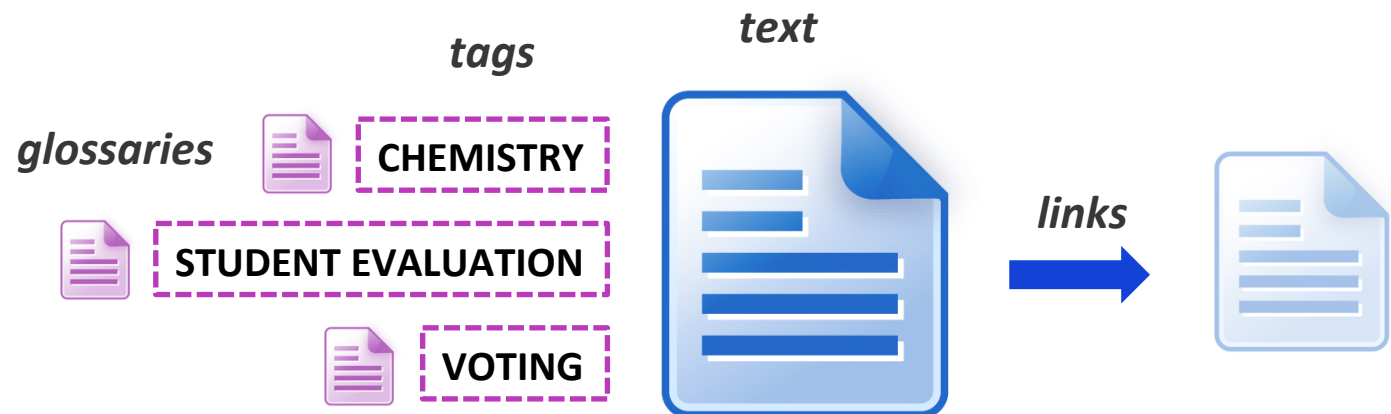
INPUT

NETWORK

NEW DOCUMENT

OUTPUT

NEW NETWORK

NETWORK DESCRIPTORS

NEW DOCUMENT DESCRIPTOR

VECTORIZER

NEAREST NEIGHBOUR

RANKING

1
2
3
4
5

THRESHOLD

PROPOSED

1
2
3
4
5

USER SELECTION

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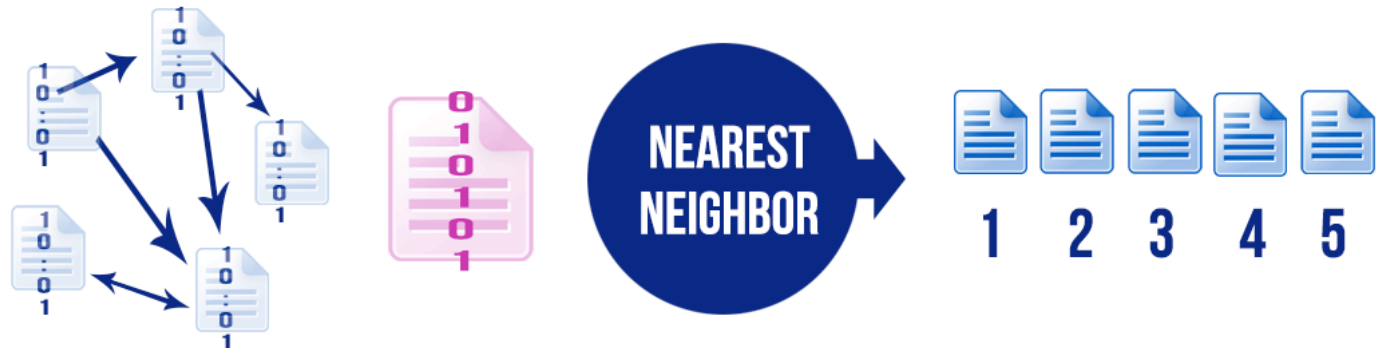
Cut off the number of returned documents based on the **difference between distances** of two consecutive ranks

CALCULATED DISTANCE

DEMONSTRATION OF OUTPUT

Performance report

%%
Performance report
%%

Average recall: 0.49723773111623578913298539646
Average precision: 0.509345794392523364485981 3083

Average recall per type
Information: 0.549745352686529157117392 4112
Question: 0.396491228070175438596491 2281
Good Practice: 0.321428571428571428571 4285715
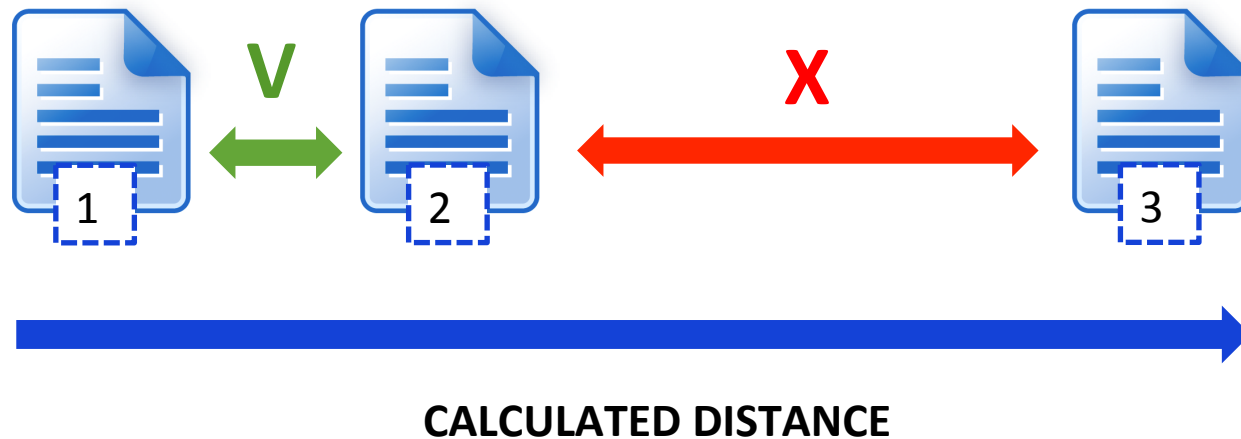Project: 0.307291666666666666666 6666666
Person: 0.582051282051282051282 0512821
Event: 0.178571428571428571428 5714286

Average precision per type
Information: 0.651960784313725490196 0784312
Question: 0.50
Good Practice: 0.375
Project: 0.625
Person: 0.393162393162393162393 1623931
Event: 0.333333333333333333333 3333333

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
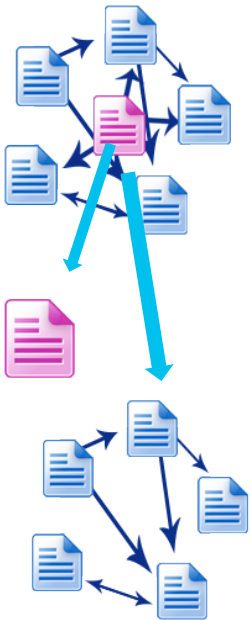
↗ **TAKE ONE OUT PRINCIPLE**

↗ **WITH THRESHOLD WE CAN MEASURE:**

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

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

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!