Finding the most important sentence

or the word in a large text

March 2018
Sezin Yaman

Introduction

This document (notebook) discusses a text summarization procedure -- finding most salient sentences -- applied to a list of questions. (Using Wolfram Language.) The data is collected from Mozilla Corporation from a large scale interview, during Jan-March 2018.

Data load

```
text = Import["/Users/yaman/Desktop/Important/ET Interviews - ET copy.txt"];
Short[text]
Length[text]
```

```
<<57 613>>
```

```
answers = StringTrim /@ StringSplit[text,
    {".", ";", "?", "“", "”", "“", ":", "??", "&", "-", "!", "’", "[", "]"}];
Short[answers];
Length[answers];
Part[answers, 1];
answers = StringReplace[answers, ___ ~~ "..." → ""];
Length[answers];
answers = StringReplace[answers, {"Content" → "",
    "Yeah" → "", "yeah" → "", "Exactly" → "", "Right" → "", "Hey" → ""}];
answers = StringReplace[answers, {"ET" → "emerging technologies"}];
answers = StringReplace[answers, NumberString → ""];
answers = ToLowerCase[answers];
```

```
%804
```

```
%794
```

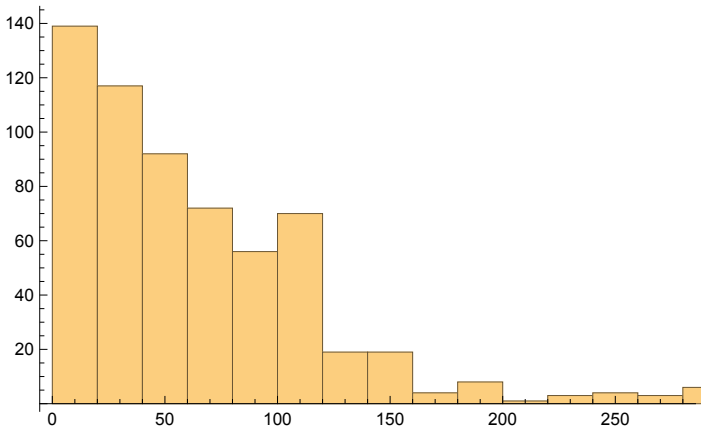
```
%784
```

```
%228

%185

%173
```

```
answers = Select[answers, StringLength[#] > 0 &];
Length[answers]
ColumnForm[RandomSample[answers, 4]]
answers = Flatten[answers];
Histogram[StringLength /@ answers]
answers = StringReplace[answers, ___ ~~ "..." → ""];
(*ColumnForm[RandomSample[answers,4]]*)
answers = Select[answers, StringLength[#] > 0 &];
answers
```



Package load

Here we load the packages [1-6]:

```
Import[
 "https://raw.githubusercontent.com/antononcube/MathematicaForPrediction/master
   /MathematicaForPredictionUtilities.m"]
Import[
 "https://raw.githubusercontent.com/antononcube/MathematicaForPrediction/master
   /DocumentTermMatrixConstruction.m"]
Import[
 "https://raw.githubusercontent.com/antononcube/MathematicaForPrediction/master
   /Misc/RSparseMatrix.m"]
Import[
 "https://raw.githubusercontent.com/antononcube/MathematicaForPrediction/master
   /MonadicProgramming/MonadicLatentSemanticAnalysis.m"]
```

Procedure theoretical description

The description in section follows [10] (and more or less copying my descriptions in [11].)

In order to make the description in this section more general we can replace the word “question” with “sentence”.

Definition of importance

Let us use the following definitions:

1. The most important sentences have the most important words, and
2. the most important words are in the most important sentences.

Calculation procedure

Starting with these definitions we do the following:

1. Convert the sentences into points in a linear vector space. Each word is an axis; each question is a point.
2. Do appropriate weighting of the sentence-word associations (IDF, TFIDF, etc.).
3. From the importance definitions we can find the eigenvectors of the representation matrix multiplied by its transpose.
4. The most important sentences will have largest coordinates in the eigenvectors.

Using graph centrality measures

We can modify steps 3 and 4 in the previous sub-section in a way that allows us to use social network analysis procedures. We can make a sentence-sentence graph and apply one or several of the graph centrality measures in WL.

(In the previous section we described the application of the Eigenvector centrality measure over the sentence-sentence similarity graph.)

Procedure application

Stop words

Here is one way to obtain English stop words, [9]:

```
stopWords =
  Complement[DictionaryLookup["*"], DeleteStopwords[DictionaryLookup["*"]]];
Length[stopWords]
RandomSample[stopWords, 12]
```

```
338
```

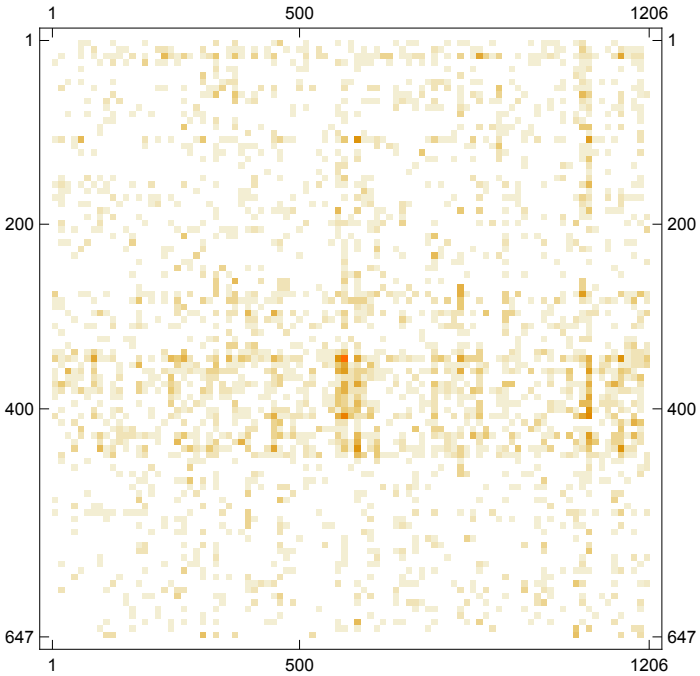
```
{side, from, and, even, whether, above,
 same-sex, the, their, custom-made, is, head-to-head}
```

Linear vector space representation

We make the following matrix of sentence-word relationships.

```
{cMat, cTerms} = DocumentTermMatrix[ToLowerCase[answers], {{}, stopWords}];
```

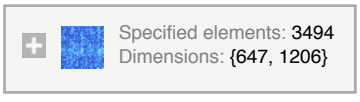
```
cMat // MatrixPlot
```



Re-weighting of the matrix entries in order to exaggerate important associations.

```
(*wcMat=WeightTerms[cMat,GlobalTermWeight["IDF",#1,#2]&,#&,#/Norm[#]&]*)
```

```
wcMat = WeightTerms[cMat]
```

```
SparseArray[  ]
```


Finding the most important sentences

Here we simulate (millions of) random walks using matrix eigen vector interpretations.

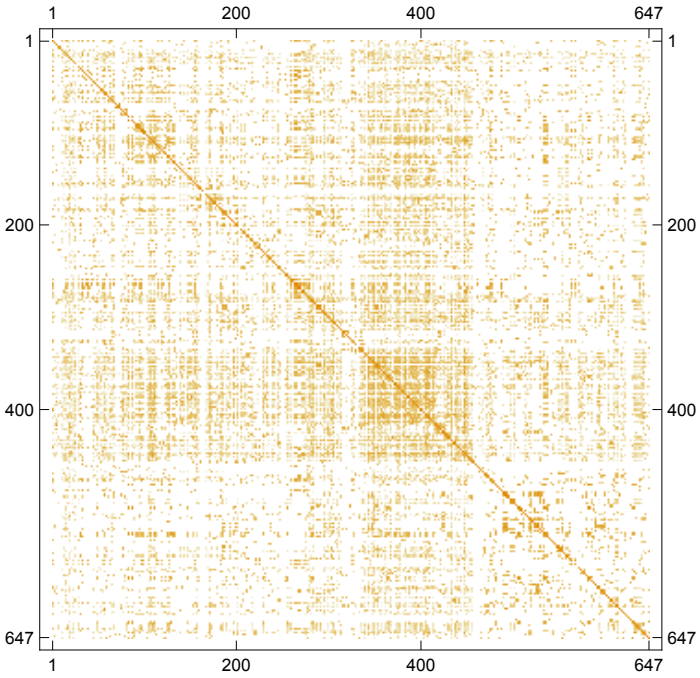
```
nSentences = 36;
```

```
wcSMat = wcMat.Transpose[wcMat]
wcSMat@"Density"
```

SparseArray [Specified elements: 33 692 Dimensions: {647, 647}]

```
0.0804856
```

```
MatrixPlot[wcSMat, AspectRatio → Automatic]
```



Removing the diagonal:

```
wcSMat = wcSMat - DiagonalMatrix[Diagonal[wcSMat]];
```

Making column stochastic:

```
wcSMat = Transpose[
  SparseArray[Map[If[Norm[#1] == 0, #1, #1 / Norm[#1]] &, Transpose[wcSMat]]]]
```

SparseArray [Specified elements: 33 180 Dimensions: {647, 647}]

Compute eigenvectors:

```
{vals, vecs} = Eigensystem[wcSMat, nSentences];
```

Display of results

Graphs

Sentence-sentence graph

Here we summarize the values of the sentence-sentence similarity matrix wcSMat (using the package [1]):

```
RecordsSummary[wcSMat@"NonzeroValues"]
```

```
  1 column 1
 Min     0.00262249
 1st Qu  0.0353249
{Median  0.0605055  }
 Mean    0.085517
 3rd Qu  0.104189
 Max     1.
```

In order to demonstrate the sentence-sentence similarity graph we are going to clip the similarity matrix values that are too small. The rest are going to be turned into 1's.

Here is the obtained graph with the top 3 most central nodes highlighted:

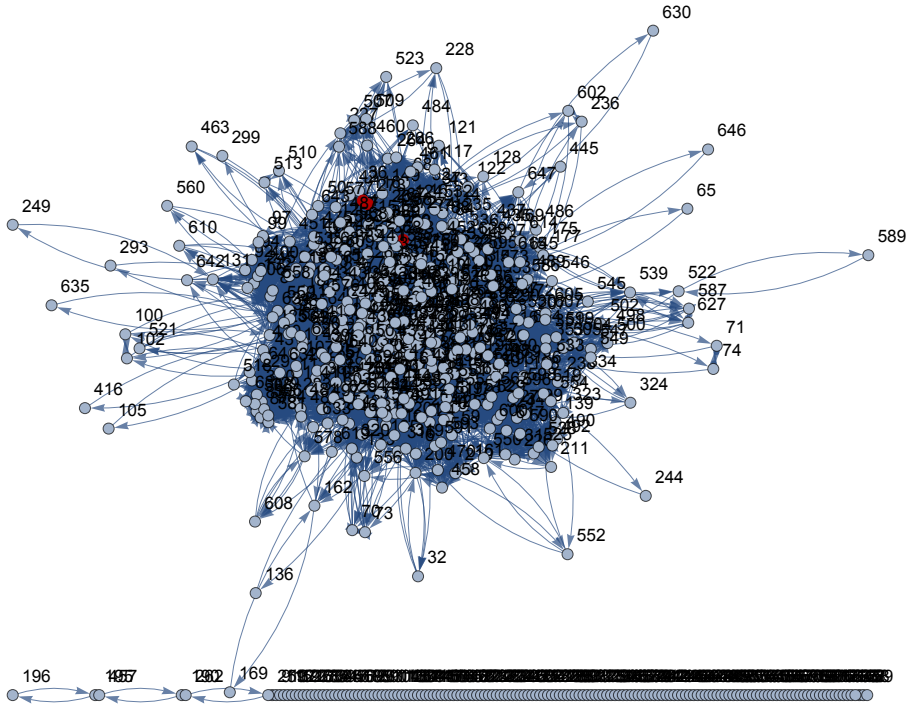
```
Block[{m = Normal@Unitize[Clip[wcSMat, {0.1, 0.3}, {0, 1}]], gr, sc},
 gr = AdjacencyGraph[m - DiagonalMatrix[Diagonal[m]],
    VertexLabels → Map[# → Tooltip[#, questions[[#]] &, Range[Length[answers]]]];
 sc = EigenvectorCentrality[gr];
 HighlightGraph[gr, Ordering[sc, -3]]
]
```

... Part: Part specification questions[[1]] is longer than depth of object.

... Part: Part specification questions[[2]] is longer than depth of object.

... Part: Part specification questions[[3]] is longer than depth of object.

... General: Further output of Part::partd will be suppressed during this calculation.



Bipartite graph

With the sentence-word matrix computed above we make a (weighed directed) bipartite graph.

Then con

```
Clear[MatrixToGraph]
MatrixToGraph[mat_?MatrixQ, col1_, col2_, opts : OptionsPattern[]] :=
  Block[{am, dims = Dimensions[mat]},
   am = SparseArray[ArrayFlatten[{{0, mat}, {Transpose[mat], 0}}]];
   AdjacencyGraph[Unitize@am, opts, DirectedEdges → True,
    VertexStyle → Join[Thread[Range[dims[[1]]] → col1],
      Thread[Range[dims[[1]] + 1, dims[[1]] + dims[[2]]] → col2]],
    GraphLayout → "BipartiteEmbedding"]
  ];
```

```
RecordsSummary[wcMat@"NonzeroValues"]
```

```
  1 column 1
  Min     0.0462707
  1st Qu  0.183906
{ Median  0.289599  }
  Mean    0.33185
  3rd Qu  0.424799
  Max     1.
```

```
Dimensions[wcMat]
```

```
{647, 1201}
```

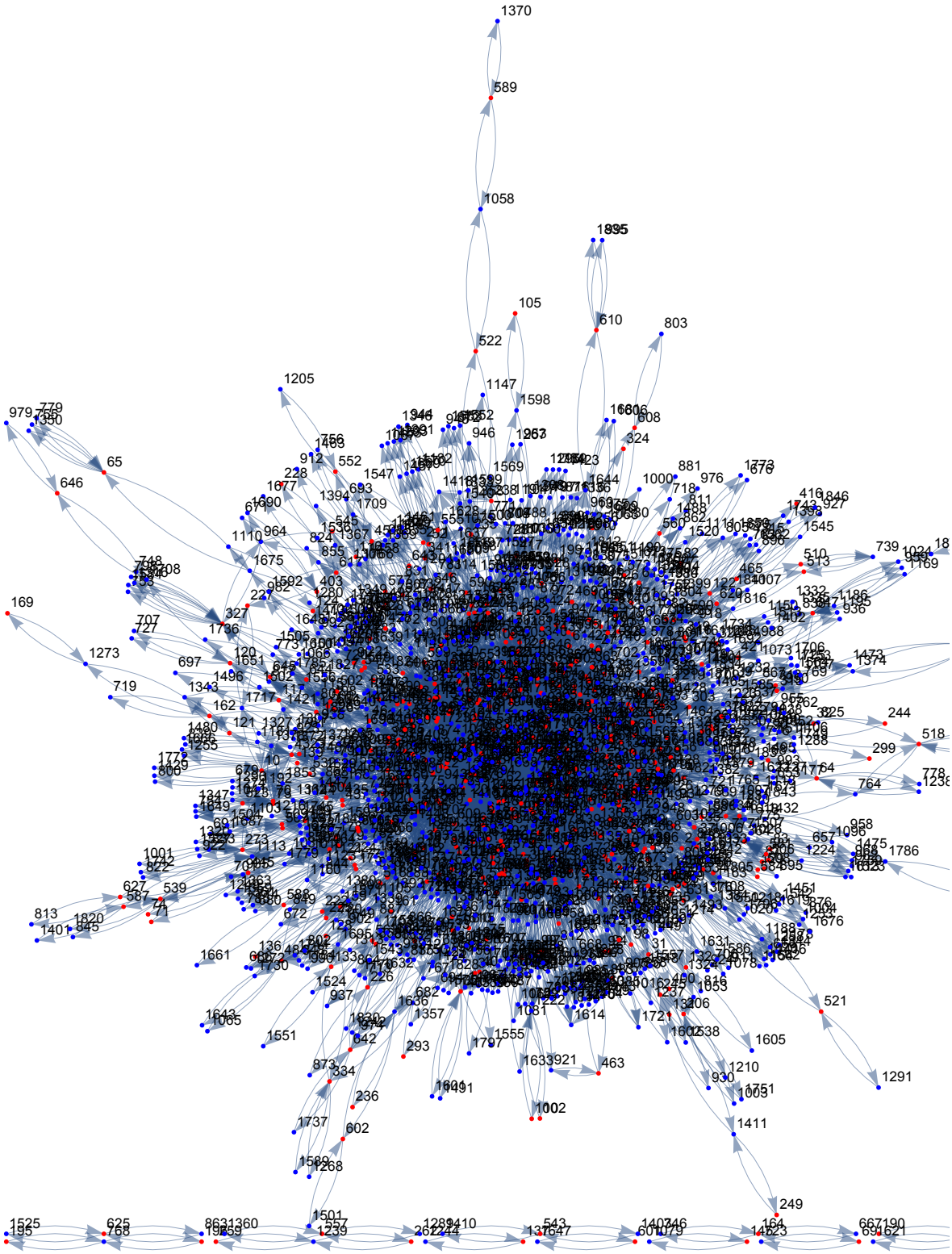
```
wcMat[[41, 305 - 114]]
```

```
0.
```

```
answers[[41]]
```

```
it
```

```
Block[{m = Unitize[wcMat], nodes = Join[answers, cTerms]},
 gr = MatrixToGraph[m, Red, Blue,
   GraphLayout → "SpringElectricalEmbedding", VertexLabels →
    Map[# → Tooltip[#, nodes[[#]]] &, Range[Length[nodes]]], ImageSize → 800]
]
```



Using LSAMon

Here we show the computations outlined and programmed above using the monad LSAMon, [1]

Stemming rules

Here we find the stemming rules for the terms found in the previous section.

```
terms = Union@StringReplace[Union@Flatten[StringSplit[ToLowerCase[answers]]],
    PunctuationCharacter → ""];
stemRules = Dispatch@Map[# → StringTrim[WordData[#, "PorterStem"]] &, terms];
```

```
%282
```

Topics extraction

```
SeedRandom[23 212]
qObj =
  LSAMonUnit[ToLowerCase /@ answers] ⟹
   LSAMonMakeDocumentTermMatrix[{}, stopWords] ⟹
   LSAMonApplyTermWeightFunctions[] ⟹
   LSAMonTopicExtraction[1, 20, 12, "MaxSteps" → 6, "PrintProfilingInfo" → True];
```

```
1 {0.060737, Null}
2 {0.083788, Null}
3 {0.080544, Null}
4 {0.08965, Null}
5 {0.080228, Null}
6 {0.085959, Null}
```

Statistical thesaurus

At this point in the object `res` we have the factors of NNMF. Using those factors we can find a statistical thesaurus for a given set of words. The following code calculates such a thesaurus, echoes it, and prints it in a tabulated form.

```
qObj ⟹ LSAMonStatisticalThesaurus({"research", "data", "mozilla", "emerging", "experiment", "user"},
    12) ⟹ LSAMonRetrieveFromContext("statisticalThesaurus") ⟹ LSAMonEchoStatisticalThesaurus();
```

word	statistical thesaurus
data	{data, little, start, deliver, certain, times, oriented, going, collect, especially, use, expect}
emerging	{emerging, technologies, maybe, siloed, kinda, disconnected, run, probably, ground, product, quickly, quarterly}
experiment	{experiment, explorers, things, kindof, big, like, audience, step, technology, careful, labelling, want}
mozilla	{mozilla, major, test, consider, market, number, quarter, high, day, abt, attraction, consumable}
research	{research, start, prototypes, basic, release, year, idea, weeks, general, getting, internally, month}
user	{user, diversed, help, understanding, nice, groups, studies, developer, testing, public, coming, set}

word	statistical thesaurus
data	{data, little, start, deliver, certain, times, oriented, going, collect, especially, use, expect}
emerging tech	{maybe, siloed, kinda, disconnected, run, probably, ground, product, quickly, quarterly}
research	{research, start, prototypes, basic, release, year, idea, weeks, general, getting, internally, month}
user	{user, diversed, help, understanding, nice, groups, studies, developer, testing, public, coming, set}

```
qObj⟹
  LSAMonStatisticalThesaurus[
   {"developer", "data", "mozilla", "manag", "experiment", "user"}, 12]⟹
  LSAMonRetrieveFromContext["statisticalThesaurus"]⟹
  LSAMonEchoStatisticalThesaurus[];
```

word	statistical thesaurus
data	{data, start, little, deliver, certain, times, especially, oriented, collect, going, use, expect}
developer	{developer, groups, coming, nice, conduct, determine, order, present, build, looks, expected, hopefully}
experiment	{experiment, explorers, things, kindof, big, like, audience, want, technology, careful, labelling, sort}
manag	{}
mozilla	{mozilla, major, test, consider, market, number, high, day, quarter, bring, rapidly, abt}
user	{user, diversed, help, understanding, nice, groups, studies, developer, testing, public, coming, set}

Sequential mining

```
t = Flatten@StringCases[ToLowerCase[answers],
   "emerging technologies" ~~ (WhitespaceCharacter ..) ~~ (x : LetterCharacter ..)]
```

```
Length[t]
```

```
22
```

```
11
```

```
11
```

```
SortBy[Tally[t], -#[[2]] &]
```

```
{{emerging technologies is, 6}, {emerging technologies and, 2},
 {emerging technologies team, 2}, {emerging technologies but, 1},
 {emerging technologies can, 1}, {emerging technologies emerging, 1},
 {emerging technologies events, 1}, {emerging technologies everyone, 1},
 {emerging technologies have, 1}, {emerging technologies monthly, 1},
 {emerging technologies people, 1}, {emerging technologies
research, 1}, {emerging technologies there, 1},
 {emerging technologies to, 1}, {emerging technologies we, 1}}
```

```
Flatten@StringCases[ToLowerCase[answers],
  "mozilla" ~~ (WhitespaceCharacter ..) ~~ (x : LetterCharacter ..) ~~
   (WhitespaceCharacter ..) ~~ (x1 : LetterCharacter ..)]
```

```
{mozilla does experiments, mozilla would do, mozilla that are,
 mozilla to work, mozilla and what, mozilla made a, mozilla has been}
```

```
t2 = Join @@ Map[Partition[StringSplit[#, WhitespaceCharacter], 2, 1] &,
    ToLowerCase[answers]];
Dimensions[
 t2]
```

```
{8697, 2}
```

```
ctMat = CrossTabulate[t2];
Dimensions[ctMat["SparseMatrix"]]
```

```
{1}
```

```
MatrixPlot[ctMat]
```

MatrixPlot: Argument CrossTabulate[{{in, et}, {et, there}, {there, could}, {could, be}, {be, an}, {an, experiment}, {experiment, of}, {of, sort}, {sort, there}, {there, is}, {is, an}, {an, audience}, {audience, of}, {of, this}, {this, technology}, {even, open}, {open, up}, {up, a}, {a, github}, {github, repository,}, ≪12≫, {to, put}, {put, it}, {it, out}, {out, there}, {here, is}, {is, an}, {an, idea,}, {idea,, we}, {we, are}, {are, gonna}, {gonna, build}, {build, sth,}, {sth,, give}, {give, it}, {it, a}, {a, few}, {few, months}, {months, and}, ≪8647≫}] at position 1 is not a matrix.

```
MatrixPlot[
 CrossTabulate[{{in, et}, {et, there}, {there, could}, {could, be}, {be, an},
   {an, experiment}, {experiment, of}, {of, sort}, {sort, there}, {there, is},
   {is, an}, {an, audience}, {audience, of}, {of, this}, {this, technology},
   {even, open}, {open, up}, ··· 8663 ···, {a, browser"}, {never, ending},
   {ending, task}, {it, is}, {is, duplicate}, {duplicate, work}, {work, to},
   {to, some}, {some, extend}, {extend, but}, {but, more}, {more, importantly},
   {importantly, i}, {i, dont}, {dont, see}, {see, the}, {the, value}}]]
```

large output | show less | show more | show all | set size limit...

```
ctRules = SortBy[Most[ArrayRules[ctMat["SparseMatrix"]]], -#[[2]] &];
```

ArrayRules: Nonrectangular array encountered.

ArrayRules: ArrayRules called with 0 arguments; 1 or 2 arguments are expected.

```
ctRules[[All, 1, 1]] = ctMat["RowNames"][[ctRules[[All, 1, 1]]]];
ctRules[[All, 1, 2]] = ctMat["ColumnNames"][[ctRules[[All, 1, 2]]]];
```

ArrayRules: ArrayRules called with 0 arguments; 1 or 2 arguments are expected.

Part: The expression ArrayRules[] cannot be used as a part specification.

ArrayRules: ArrayRules called with 0 arguments; 1 or 2 arguments are expected.

Part: The expression ArrayRules[] cannot be used as a part specification.

```
Take[ctRules, 20]
```

Take: Cannot take positions 1 through 20 in ArrayRules[].

```
Take[ArrayRules[], 20]
```

Extracted topics

Let us tabulate the topics found `LSAMonTopicExtraction` above:

```
qObj⟹
  LSAMonEchoTopicsTable["NumberOfTableColumns" → 8, "NumberOfTerms" → 12,
   "MagnificationFactor" → 0.8, Appearance → "Horizontal"];
```

1		2		3		4		5	
1.000	actually	1.000	connects	1.000	social	1.000	unclear	1.000	non
0.386	visible	0.850	facebook	0.874	useful	0.803	sometimes	1.000	middle
0.299	convenient	0.618	thing	0.632	know	0.783	greg	0.810	ground
0.213	make	0.268	services	0.602	really	0.702	decisions	0.721	probabl
0.140	impact	0.111	party	0.576	stay	0.398	making	0.539	product
0.127	register	0.089	customers	0.484	fail	0.343	outside	0.066	quarter
0.125	wrong	0.080	wanna	0.470	gonna	0.312	common	0.053	frustra
0.122	framing	0.079	users	0.428	long	0.206	different	0.045	trends
0.120	figure	0.076	auto	0.411	group	0.200	street	0.045	numbers
0.114	identity	0.069	right	0.397	succeed	0.194	deep	0.045	meant
0.111	users	0.068	logged	0.389	understanding	0.179	trust	0.045	diagnos
0.095	top	0.068	discovery	0.388	viable	0.170	voice	0.041	use
9		10		11		12		13	
1.000	bit	1.000	experiment	1.000	company	1.000	testing	1.000	develop
0.749	disconnected	0.697	explorers	0.273	want	0.635	user	0.353	process
0.537	maybe	0.424	things	0.271	impact	0.531	public	0.200	like
0.410	siloed	0.294	like	0.210	old	0.393	testpilot	0.168	just
0.054	quickly	0.218	kindof	0.198	channels	0.391	set	0.156	frame
0.053	news	0.166	want	0.185	feature	0.371	works	0.140	rust
0.048	little	0.160	big	0.177	simple	0.336	product	0.107	point
0.047	source	0.138	audience	0.177	money	0.282	good	0.097	student
0.046	thing	0.113	technology	0.177	cost	0.280	news	0.086	party
0.046	public	0.110	step	0.177	choice	0.277	work	0.078	sets
0.045	hear	0.104	labelling	0.177	buy	0.271	end	0.061	end
0.043	kinda	0.104	careful	0.177	topic	0.266	firefox	0.059	sound
17		18		19		20			
1.000	making	1.000	job	1.000	studies	1.000	need		
0.830	decision	0.992	better	0.716	diversed	0.669	feedback		
0.387	fed	0.810	showing	0.663	user	0.356	solicit		
0.378	communicate	0.617	building	0.545	understanding	0.190	driven		
0.377	managers	0.518	process	0.535	help	0.182	feel		
0.289	experiments	0.482	kind	0.276	groups	0.177	internal		
0.272	process	0.454	short	0.275	nice	0.116	visible		
0.171	lot	0.454	foundational	0.236	developer	0.096	experimentation		
0.146	evidence	0.400	tangible	0.173	coming	0.092	anxiety		
0.127	visible	0.399	siloed	0.143	conduct	0.079	point		
0.119	‘cause	0.396	term	0.139	determine	0.074	engagement		
0.116	anxiety	0.382	tech	0.106	present	0.071	community		

Most important sentences

Using sentence-word bipartite graph

Using the default construction of bipartite graph

```
res =
  qObj⟹
    LSAMonMostImportantTexts[12, "CentralityFunction" → EigenvectorCentrality];
GridTableForm[res⟹LSAMonTakeValue, TableHeadings → {"score", "index", "text"}]
```

#	score	index	text
1	0.0148074	1262	like
2	0.0132665	1235	know
3	0.00916957	1728	think

4	0.00850892	361	at a certain level of traction there is a very rapidly emerging surface area of problems that suddenly become possible to start to analyze in a more concrete in a peculiar way for example everything at the top of the user's funnel experience so like how they sign up whether they get their headset on correctly or whether they get sick in the first thirty seconds a lot of these really first order concerns about acquisition and engagement of users i think, can rapidly become analytical data base at least in part in decision making and so it's important that you kind of be prepared for when that happens you know i think you know you have to have you know thousands of users over some reasonable period of time to start making sense of things but in general you know that becomes an important transition point for a product i think
5	0.0085056	1727	things
6	0.0078182	447	no users more like you know the two came in this week from much and out an early stage i feel like maybe our early group of users are going to be pretty small i mean maybe it'll blow up but i doubt it it doesn't have the characteristics that i would expect anything to blow up so early on i feel like we're going to end up with having a much more personal relationship with all of our developer users and we're going to be talking with them and we're going to be jumping into their application helping things like us once it gets beyond that i you know one bridge vision of the future is that there might be thirty applications using our underlying technology and each of those applications have their own community and those communities will have to different values and so the services the underlying services will maybe need to provide different things to serve different types of communities and in that situation i can imagine using user research either from end users of the already users like you know or we get so many developers we start running like project we start running events where you come you learn about our project and you build your own little thing and we want to find out ok what was difficult what was easy and where are what are we playing what do we need to document better

7	0.00738991	344	technical feasibility mode so basically just internally building a prototype to determine if we could actually deliver a good experience so there wasn't really any user and end user oriented data collection at that point and then you know the first month of this year we're effectively kind of planning and trying to scope our plans for the year and so on so you know we're just kind of hitting the initial couple weeks here of kind of focus product development and so to that end you know our near term plan is to do some type of basic lightweight design process that those take into account some kind of user studies on what that might look like is is building very very basic ruffy or prototypes and d prototypes that for and you know getting feedback from kind of general audience
8	0.00714739	349	so i mean one thing that's kind of unique about our team is all of us of work together you know number of years and so we've also worked on some where products before and so we have quite deep track record of experience that that encompasses both subjective knowledge about what users have had good and bad experiences with but we also have a good amount of data that you know to some degree in our heads could may or may not to extrapolate forward to our current efforts, so some of the things like that are in the back of our minds constantly are things like expected user session times and the expected device support and the you know the breath of different platforms that people are going to be on and all those kinds of things
9	0.00704217	384	in addition to that there is actually a large surface area of functionality that is extremely potentially valuable to users that rely upon similar types of data collection for example if i want to record myself and play myself back as an avatar, you know what what can of worms does that open now or that was the same as the experiment i don't really know but you know it's a it's a whole kind of you know new area that there's not a whole lot of existing cultural things to draw on because you're now basically be able to effectively simulate the person's body like which never think james so
10	0.00643965	1514	really
11	0.00622267	1230	kind

12	0.0060429	393	and so it's i think it's more of a style thing though it doesn't it doesn't mean that there's only one way to do things right in general like i think particularly if you're trying to uncover you know really really strong immutable truths like a more kind of traditional research oriented approach makes a lot of sense

Using sentence-sentence graph and removing (self-)loops

```
res =
  qObj ⟹ LSAMonMakeGraph["Type" → "DocumentDocument"] ⟹
   LSAMonMostImportantTexts[12, "CentralityFunction" → EigenvectorCentrality];
GridTableForm[res ⟹ LSAMonTakeValue, TableHeadings → {"score", "index", "text"}]
```

... DiagonalMatrix: Input matrix contains an infinite entry.

#	score	index	text
1	0.0015456	647	it is duplicate work to some extend but more importantly i dont see the value
2	0.0015456	433	well i think that we all have about one hundred things we can think of working on the way we decided to prioritize them are grouping them in chunks that are useful when delivered together and we decided on a set of features that are the most important to deliver first and we kind of broke it up on a quarterly basis and thought ok what what can we expect to complete in the first quarter and these are the things that we are most clearly sure that this is going to be something we want to work on then second quarter it's like
3	0.0015456	432	
4	0.0015456	431	no but i think that's because we haven't really gotten that first big step
5	0.0015456	430	no but i think that's because we haven't really gotten that first big step
6	0.0015456	429	they are having a problem with theirs or the other person was just asking hey like asking a general question to check it's understanding of what we're doing
7	0.0015456	428	two people that i can remember this last week of comment that
8	0.0015456	427	already even though we haven't done any official release there are people our site tell us in what we've done and other and asking questions online of people
9	0.0015456	426	so we have a slack channel so that they care talk with us we have our code online and we haven't done an official release to them yet but the plan is that once we do people will start using the library and then any issues that they run into they can file on the on github where you know the code is or through slack channel and we can just chat with
10	0.0015456	425	i'm going to ask about the feedback mechanisms how do you get information from and users
11	0.0015456	424	
12	0.0015456	423	i'm going to ask about the feedback mechanisms how do you get information from and users

Using sentence-sentence graph with (self-)loops

```
res =
  qObj ⟹ LSAMonMakeGraph["Type" → "DocumentDocument", "RemoveLoops" → False] ⟹
   LSAMonMostImportantTexts[12, "CentralityFunction" → EigenvectorCentrality];

GridTableForm[res ⟹ LSAMonTakeValue, TableHeadings → {"score", "index", "text"}]
```

#	score	index	text
1	0.00728913	361	at a certain level of traction there is a very rapidly emerging surface area of problems that suddenly become possible to start to analyze in a more concrete in a peculiar way for example everything at the top of the user's funnel experience so like

Printed by Wolfram Mathematica Student Edition

2	0.00703024	384	in addition to that there is actually a large surface area of functionality that is extremely potentially valuable to users that rely upon similar types of data collection for example if i want to record myself and play myself back as an avatar, you know what what can of worms does that open now or that was the same as the experiment i don't really know but you know it's a it's a whole kind of you know new area that there's not a whole lot of existing cultural things to draw on because you're now basically be able to effectively simulate the person's body like which never think james so
3	0.00702261	349	so i mean one thing that's kind of unique about our team is all of us of work together you know number of years and so we've also worked on some where products before and so we have quite deep track record of experience that that encompasses both subjective knowledge about what users have had good and bad experiences with but we also have a good amount of data that you know to some degree in our heads could may or may not to extrapolate forward to our current efforts, so some of the things like that are in the back of our minds constantly are things like expected user session times and the expected device support and the you know the breath of different platforms that people are going to be on and all those kinds of things
4	0.00698493	393	and so it's i think it's more of a style thing though it doesn't it doesn't mean that there's only one way to do things right in general like i think particularly if you're trying to uncover you know really really strong immutable truths like a more kind of traditional research oriented approach makes a lot of sense
5	0.00690315	406	i think everyone and i think the i think the best forms of management you know are are really basically about providing like rather high level of guidance and trusting the people on the ground building stuff to make the most of the tactical almost all the tactical decisions at a very large number of strategic decisions and to that end i think everyone has done a really good job all up and down you know i think

6	0.00689454	447	no users more like you know the two came in this week from much and out an early stage i feel like maybe our early group of users are going to be pretty small i mean maybe it'll blow up but i doubt it it doesn't have the characteristics that i would expect anything to blow up so early on i feel like we're going to end up with having a much more personal relationship with all of our developer users and we're going to be talking with them and we're going to be jumping into their application helping things like us once it gets beyond that i you know one bridge vision of the future is that there might be thirty applications using our underlying technology and each of those applications have their own community and those communities will have to different values and so the services the underlying services will maybe need to provide different things to serve different types of communities and in that situation i can imagine using user research either from end users of the already users like you know or we get so many developers we start running like project we start running events where you come you learn about our project and you build your own little thing and we want to find out ok what was difficult what was easy and where are what are we playing what do we need to document better
7	0.00678743	413	you know i think i think your primary role as a as a manager i hate to use the term manager i would kind of consider it more like a leadership thing and i think that's kind of with arnold to any time or chart but in general as any kind of leader you know your major your major responsibility is to basically ensure people's well being and feeling fulfillment in their day to day and so on to that and i think mozilla has been extremely like great in terms of people understanding that that's really incredible also you know i think this is rea
8	0.00678078	347	so you know we will be in the fortunate position of having a user facing product hopefully within months and so that i think will be a rich you know opportunity for us to to collect all kinds of feedback from you know anecdotal user direct feedback to data on user behavior and things like this in so far as we're able to so

9	0.00668943	439	trying to think i think that necessarily there will be some throwaway work as we as we kind of feel out the ecosystem and we're also using tools that are we are much less familiar with than what we used previously and like a big investigation we did this month which is time consuming but that's maybe a big maybe necessary was exploring a larger variety of possible tools we have use which you know once you decide like ok these are the tools we're going to use then you could imagine well all that work you did learning the other tools that sort of you have this kind of useless but it's sort of
10	0.00664436	344	technical feasibility mode so basically just internally building a prototype to determine if we could actually deliver a good experience so there wasn't really any user and end user oriented data collection at that point and then you know the first month of this year we're effectively kind of planning and trying to scope our plans for the year and so on so you know we're just kind of hitting the initial couple weeks here of kind of focus product development and so to that end you know our near term plan is to do some type of basic lightweight design process that those take into account some kind of user studies on what that might look like is is building very very basic ruffy or prototypes and d prototypes that for and you know getting feedback from kind of general audience
11	0.00643387	405	i think you know validating your learning as early as possible and having a certain level of inclusiveness in both visibility into what's happening and involvement and so far as people who would be considered stakeholders feel like they have involvement i think that can make those types of outcomes a little more digestible
12	0.00636457	378	effectively pond on the implications, because we were typically at smaller organizations where like there was plenty of room for us to adjust later if as we scaled up to the user base but so i think we'll have to do a little bit more upfront thinking about that and it's a good thing to do and i know that you know this is alrready done in firefox to some degree

```
res =
  qObj ⟹ LSAMonMakeGraph["Type" → "DocumentDocument", "RemoveLoops" → False] ⟹
   LSAMonMostImportantTexts[20, "CentralityFunction" → EigenvectorCentrality];
```

```
GridTableForm[res⟹LSAMonTakeValue, TableHeadings→{"score", "index", "text"}]
```

#	score	index	text
1	0.00728913	361	at a certain level of traction there is a very rapidly emerging surface area of problems that suddenly become possible to start to analyze in a more concrete in a peculiar way for example everything at the top of the user's funnel experience so like how they sign up whether they get their headset on correctly or whether they get sick in the first thirty seconds a lot of these really first order concerns about acquisition and engagement of users i think, can rapidly become analytical data base at least in part in decision making and so it's important that you kind of be prepared for when that happens you know i think you know you have to have you know thousands of users over some reasonable period of time to start making sense of things but in general you know that becomes an important transition point for a product i think
2	0.00703024	384	in addition to that there is actually a large surface area of functionality that is extremely potentially valuable to users that rely upon similar types of data collection for example if i want to record myself and play myself back as an avatar, you know what what can of worms does that open now or that was the same as the experiment i don't really know but you know it's a it's a whole kind of you know new area that there's not a whole lot of existing cultural things to draw on because you're now basically be able to effectively simulate the person's body like which never think james so

3	0.00702261	349	so i mean one thing that's kind of unique about our team is all of us of work together you know number of years and so we've also worked on some where products before and so we have quite deep track record of experience that that encompasses both subjective knowledge about what users have had good and bad experiences with but we also have a good amount of data that you know to some degree in our heads could may or may not to extrapolate forward to our current efforts, so some of the things like that are in the back of our minds constantly are things like expected user session times and the expected device support and the you know the breath of different platforms that people are going to be on and all those kinds of things
4	0.00698493	393	and so it's i think it's more of a style thing though it doesn't it doesn't mean that there's only one way to do things right in general like i think particularly if you're trying to uncover you know really really strong immutable truths like a more kind of traditional research oriented approach makes a lot of sense
5	0.00690315	406	i think everyone and i think the i think the best forms of management you know are are really basically about providing like rather high level of guidance and trusting the people on the ground building stuff to make the most of the tactical almost all the tactical decisions at a very large number of strategic decisions and to that end i think everyone has done a really good job all up and down you know i think

6	0.00689454	447	no users more like you know the two came in this week from much and out an early stage i feel like maybe our early group of users are going to be pretty small i mean maybe it'll blow up but i doubt it it doesn't have the characteristics that i would expect anything to blow up so early on i feel like we're going to end up with having a much more personal relationship with all of our developer users and we're going to be talking with them and we're going to be jumping into their application helping things like us once it gets beyond that i you know one bridge vision of the future is that there might be thirty applications using our underlying technology and each of those applications have their own community and those communities will have to different values and so the services the underlying services will maybe need to provide different things to serve different types of communities and in that situation i can imagine using user research either from end users of the already users like you know or we get so many developers we start running like project we start running events where you come you learn about our project and you build your own little thing and we want to find out ok what was difficult what was easy and where are what are we playing what do we need to document better
7	0.00678743	413	you know i think i think your primary role as a as a manager i hate to use the term manager i would kind of consider it more like a leadership thing and i think that's kind of with arnold to any time or chart but in general as any kind of leader you know your major your major responsibility is to basically ensure people's well being and feeling fulfillment in their day to day and so on to that and i think mozilla has been extremely like great in terms of people understanding that that's really incredible also you know i think this is rea
8	0.00678078	347	so you know we will be in the fortunate position of having a user facing product hopefully within months and so that i think will be a rich you know opportunity for us to to collect all kinds of feedback from you know anecdotal user direct feedback to data on user behavior and things like this in so far as we're able to so

9	0.00668943	439	trying to think i think that necessarily there will be some throwaway work as we as we kind of feel out the ecosystem and we're also using tools that are we are much less familiar with than what we used previously and like a big investigation we did this month which is time consuming but that's maybe a big maybe necessary was exploring a larger variety of possible tools we have use which you know once you decide like ok these are the tools we're going to use then you could imagine well all that work you did learning the other tools that sort of you have this kind of useless but it's sort of
10	0.00664436	344	technical feasibility mode so basically just internally building a prototype to determine if we could actually deliver a good experience so there wasn't really any user and end user oriented data collection at that point and then you know the first month of this year we're effectively kind of planning and trying to scope our plans for the year and so on so you know we're just kind of hitting the initial couple weeks here of kind of focus product development and so to that end you know our near term plan is to do some type of basic lightweight design process that those take into account some kind of user studies on what that might look like is is building very very basic ruffy or prototypes and d prototypes that for and you know getting feedback from kind of general audience
11	0.00643387	405	i think you know validating your learning as early as possible and having a certain level of inclusiveness in both visibility into what's happening and involvement and so far as people who would be considered stakeholders feel like they have involvement i think that can make those types of outcomes a little more digestible
12	0.00636457	378	effectively pond on the implications, because we were typically at smaller organizations where like there was plenty of room for us to adjust later if as we scaled up to the user base but so i think we'll have to do a little bit more upfront thinking about that and it's a good thing to do and i know that you know this is alrready done in firefox to some degree

13	0.00634515	106	you know with her like we're there we done in the past play pointer to academic research that's been done but also help her with the logistics like here's how i would recruit participants like maybe you should think about this methodology maybe you should use this approach to your study etc so like that was a good conversation
14	0.00627022	367	liveliness like effectively when the system has active users using it daily if you have good visibility on that system and good monitoring in metrics and a good operations culture around the system your engineering processes change in a way that is i think extremely healthy and important and it makes it so that you have a much more higher level of confidence in the changes your making because you have live validation that things are continuing to work whereas when you're in this pre
15	0.00625482	127	i feel like i know the et people pretty well, and i like them and i respect them and i think they
16	0.00623096	377	i think there's definitely some place there that we can get a lot of really useful understanding without crossing the lines but you know having worked at start up like environments in the past we've generally erred on the side of let's collect the data, as long as it's motivated for some useful analysis purpose and then
17	0.00615038	354	so i mean we've kind of been going through the process since the beginning of the year of just you know trying to document our thinking and you know that's communicated through pretty traditional things like word documents and you know good decks and things like this
18	0.00607548	433	well i think that we all have about one hundred things we can think of working on the way we decided to prioritize them are grouping them in chunks that are useful when delivered together and we decided on a set of features that are the most important to deliver first and we kind of broke it up on a quarterly basis and thought ok what what can we expect to complete in the first quarter and these are the things that we are most clearly sure that this is going to be something we want to work on then second quarter it's like

19	0.00604925	157	so, for instance, if they really wanna work on like a connective device, sort of cloud based connected device, i think one of the things they should really be thinking about is what are the assets that firefox can bring, right
20	0.00603412	452	if we have a product like that that reaches end users and i suspect there will be plenty of opportunity to include user research into that project i'm also curious you know myriam does more forward thinking around what are the markets of the future like what are the means of people not today but in two three five years

References

Packages

[1] Anton Antonov, MathematicaForPrediction utilities *Mathematica* package, (2014), MathematicaForPrediction at GitHub.

[2] Anton Antonov, Implementation of document-term matrix construction and re-weighting functions in Mathematica, (2013), MathematicaForPrediction at GitHub.

[3] Anton Antonov, Implementation of the Non-Negative Matrix Factorization algorithm in Mathematica, (2013), MathematicaForPrediction at GitHub.

[4] Anton Antonov, Implementation of one dimensional outlier identifying algorithms in Mathematica, (2013), MathematicaForPrediction at GitHub.

[5] Anton Antonov, Monadic latent semantic analysis Mathematica package, (2017), MathematicaForPrediction at GitHub.

[6] Anton Antonov, RSparseMatrix Mathematica package, (2015), MathematicaForPrediction at GitHub.

Articles

[7] Wikipedia entry, Latent Semantic Analysis.

[8] Wikipedia entry, Bag-of-words model.

[9] Wikipedia entry, Stop words.

[10] Lars Elden, Matrix Methods in Data Mining and Pattern Recognition, 2007, SIAM. See Chapter 13, "Automatic Key Word and Key Sentence Extraction".

[11] Anton Antonov, Answer of the question “To sermon or not to sermon”, (2016), MathematicaStackExchange Meta,
URL: https://mathematica.meta.stackexchange.com/a/1858/34008 .

[12] Anton Antonov, “Statements saliency in podcasts”, (2016), MathematicaVsR at GitHub.
URL: https://github.com/antononcube/MathematicaVsR/tree/master/Projects/State-

mentsSaliencyInPodcasts .

[13] Anton Antonov, “The Great conversation in USA presidential speeches”, (2017), MathematicaForPrediction at WordPress blog.

URL: https://mathematicaforprediction.wordpress.com/2017/12/24/the-great-conversation-in-usa-presidential-speeches/ .