Exploration of Large Document Sets

Mark Cieliebak

13.1.2021

Exploration of Large Document Sets:
What is the underlying data?

- 500…1'000'000 documents
- Homogeneous or heterogeneous documents
- Arbitrary data sources, e.g. news, scientific articles, project reports etc.
- Arbitrary data formats, e.g. txt, pdf, word, OCR output etc.
- Plain text or with meta data (author, date, source, ID etc.)

Exploration of Large Document Sets:
What are the potential goals?

1. Find documents for specific keywords/topics (search)
2. Find the most relevant document (search, filter)
3. Extract structured information (e.g. NER)
4. Get an overview
5. Find "something interesting"

Sample Applications for Large Document Set Exploration

Historical Analysis

- Goal: understand development of a specific topic over time
- 500-50k news articles

Scientific Research:

- Goal: quick overview of a research field
- 1'000-50k papers for a specific topic

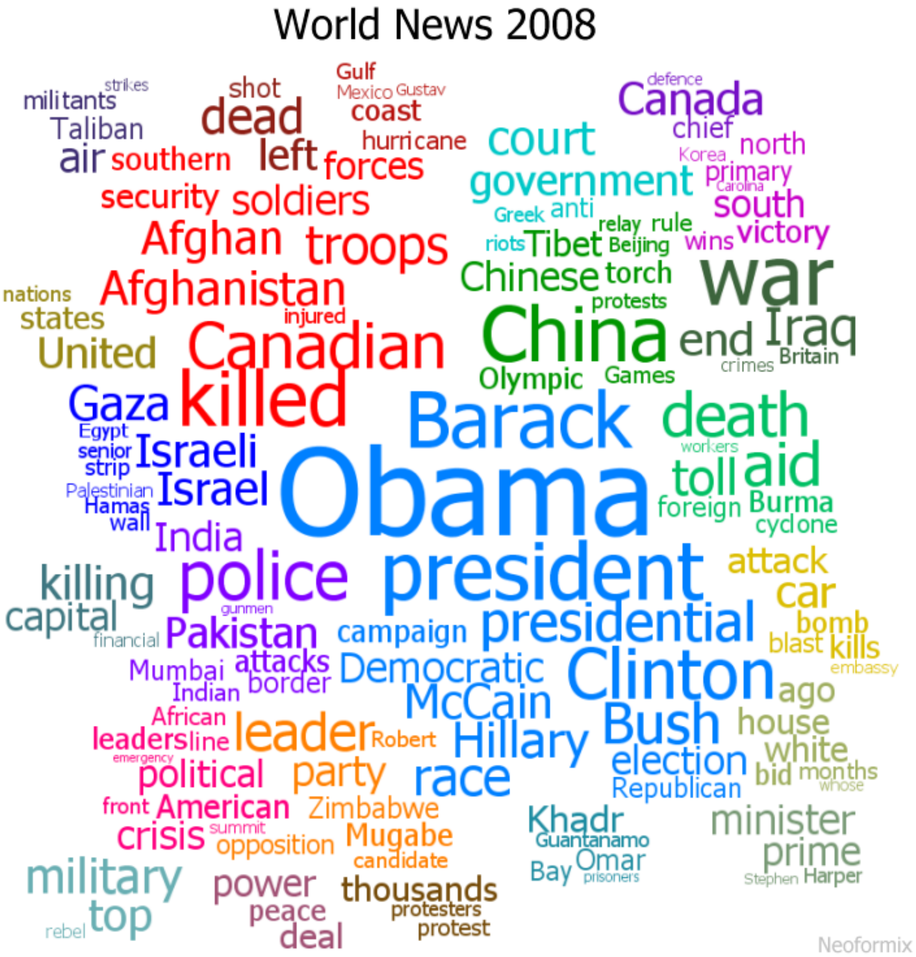
Callenge a Patent:

- Goal: for a given patent, find the most similar patents
- Size: 14 million patents worldwide, 1-2k after pre-filtering

Police Forensics:

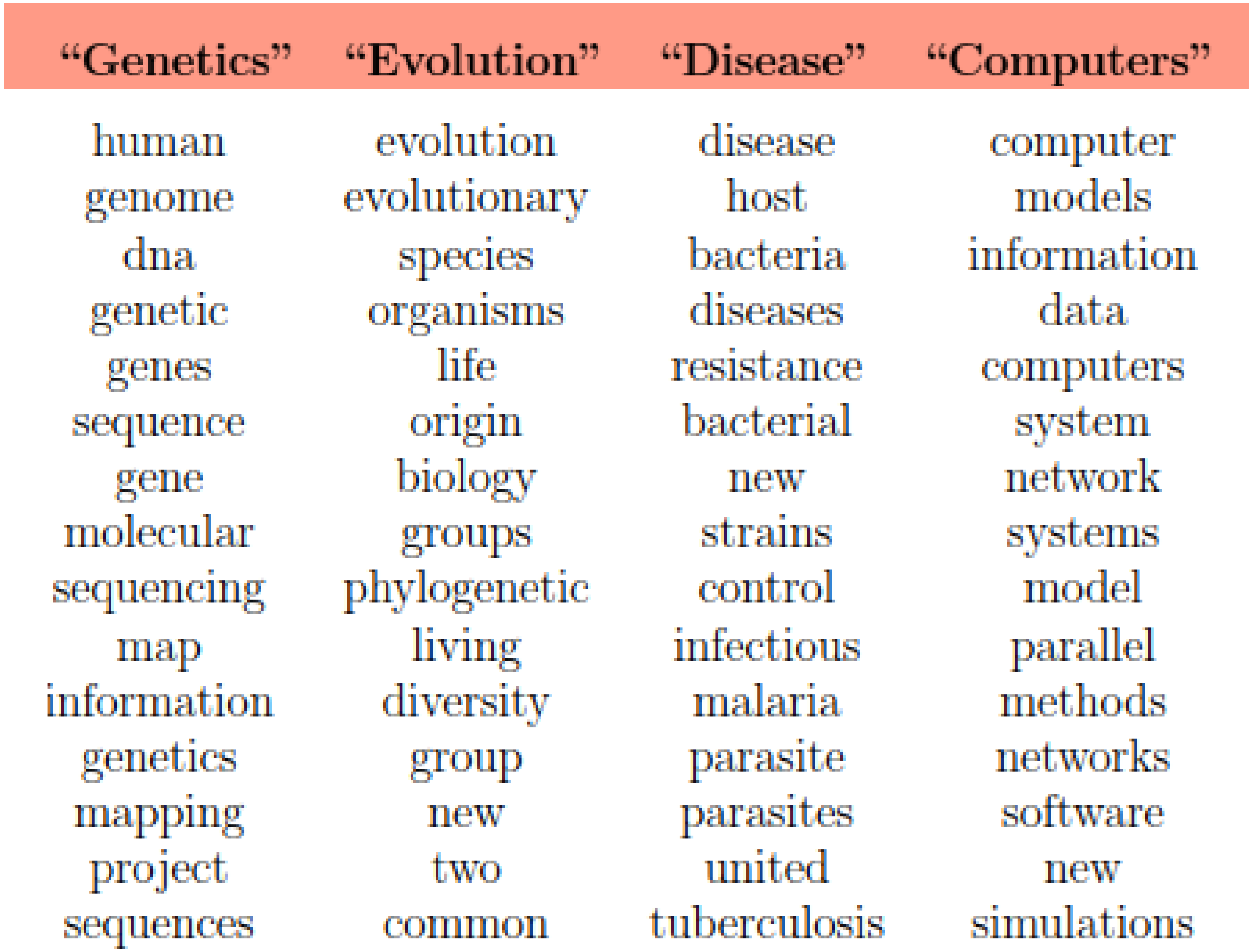
- Goal: find "interesting" documents for a criminal case on a harddisk
- Size: 100k-500k docs

Keyword Extraction can find important terms in single documents or document (sub-)sets



Methods:
- Word Frequency
- TF-IDF
- RAKE
- Textrank

Topic Modeling can identify topics in large document sets

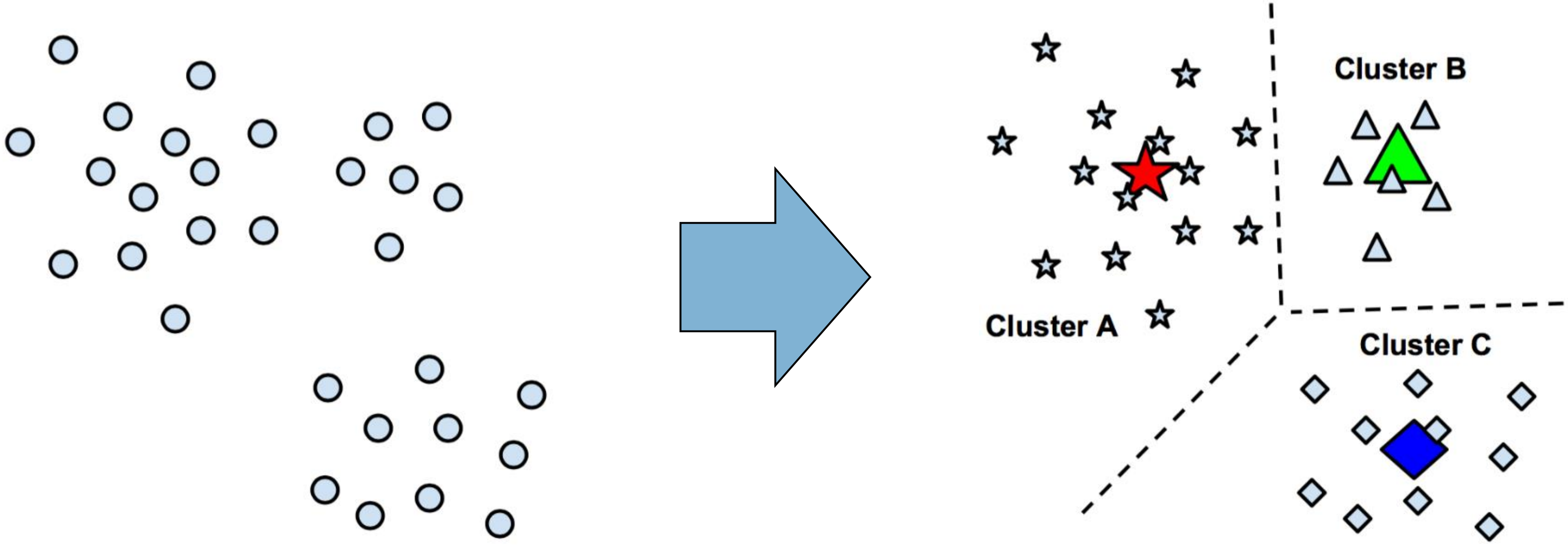


"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Methods:
- LDA
- LSA

100-topic LDA model to 17,000 articles from the journal "Science"

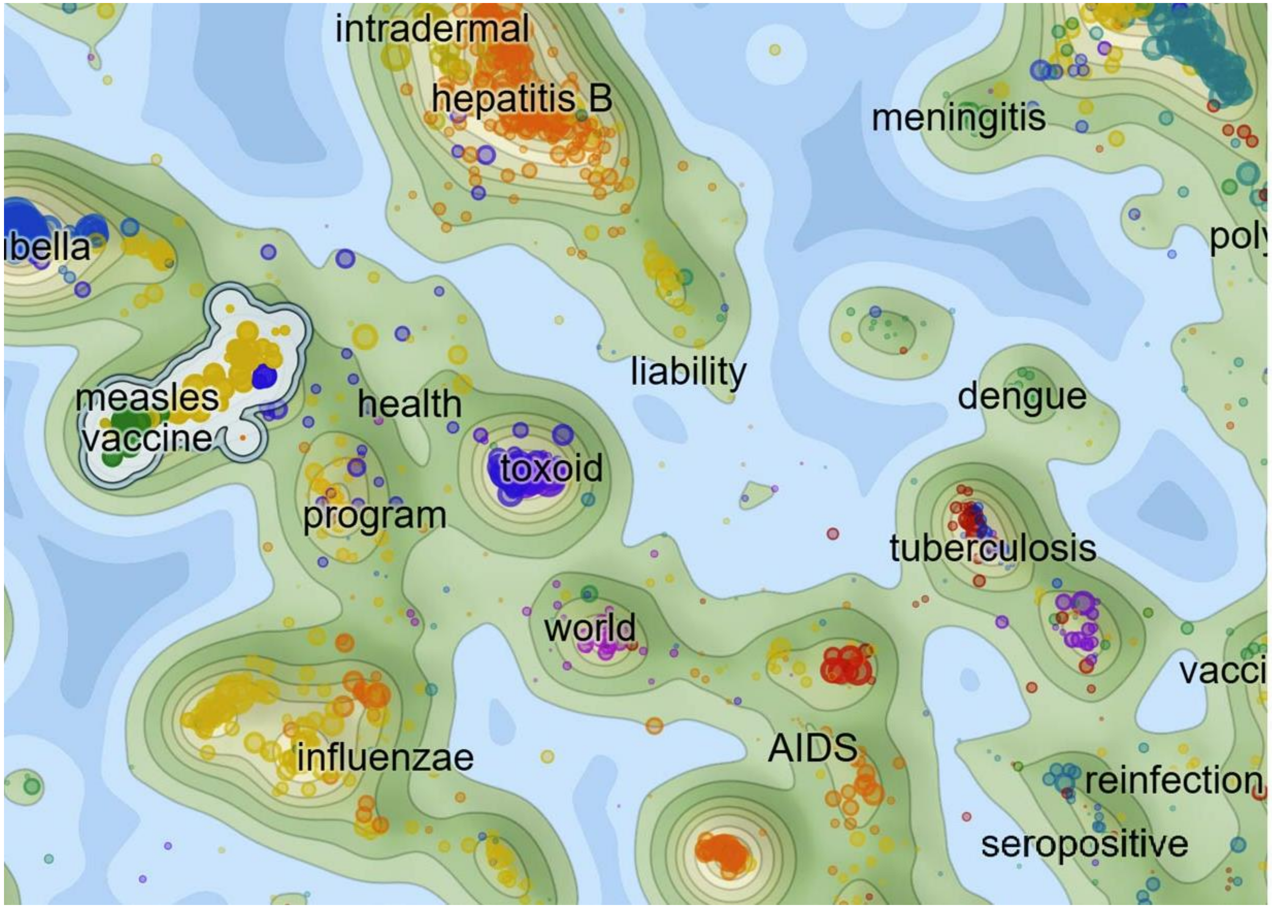
Clustering can group similar documents



Methods:

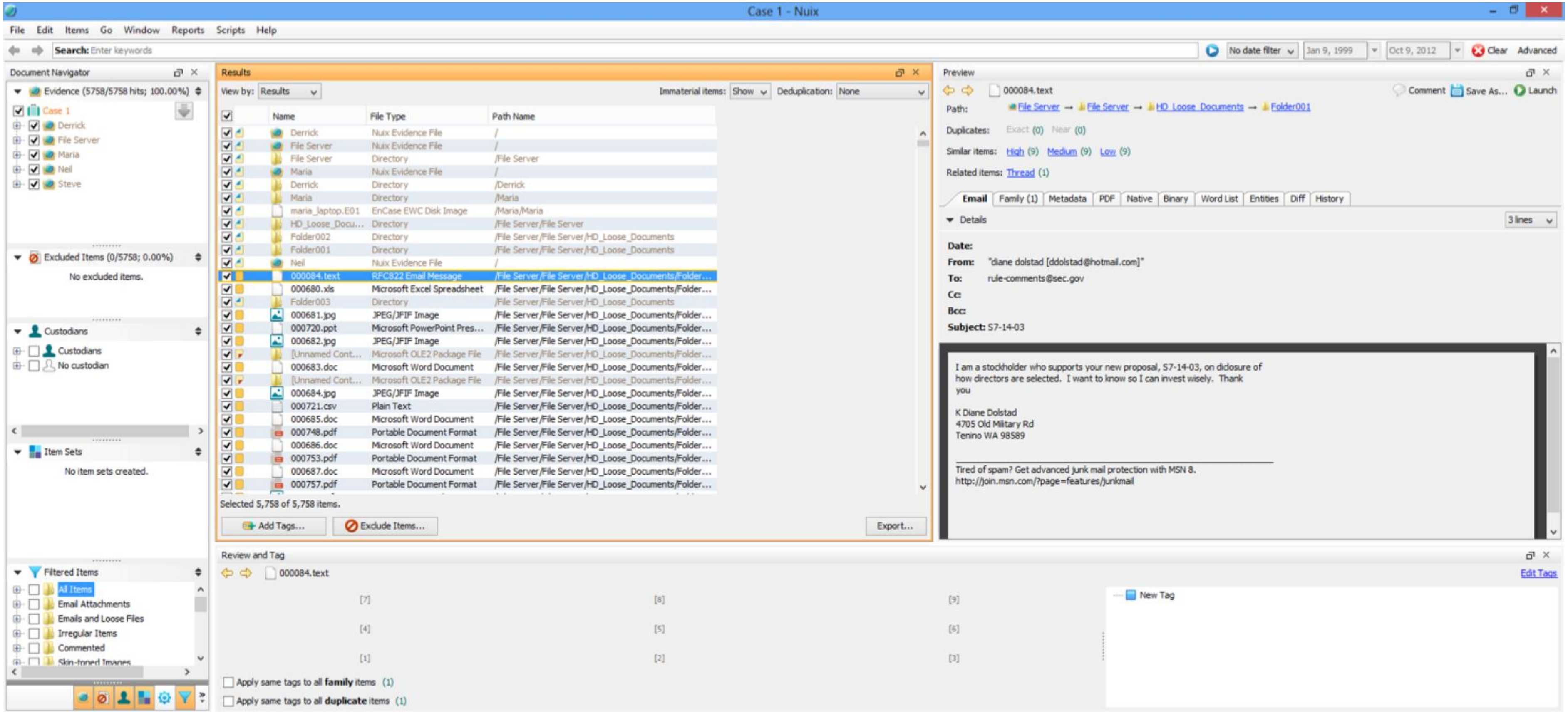
- K-Means
- Agglomerative Clustering
- DB-SCAN
- Gaussian Mixture Models

Visualization of clustering scientific article; distance is "textual similarity" of the abstracts



https://carrotsearch.com/lingo4g

There exist tools to explore a computer hard drive interactively



☑	Name	File Type	Path Name
☑	Derrick	Nuix Evidence File	/
☑	File Server	Nuix Evidence File	/
☑	File Server	Directory	/File Server
☑	Maria	Nuix Evidence File	/
☑	Derrick	Directory	/Derrick
☑	Maria	Directory	/Maria
☑	maria_laptop.E01	EnCase EWC Disk Image	/Maria/Maria
☑	HD_Loose_Docu...	Directory	/File Server/File Server
☑	Folder002	Directory	/File Server/File Server/HD_Loose_Documents
☑	Folder001	Directory	/File Server/File Server/HD_Loose_Documents
☑	Neil	Nuix Evidence File	/
☑	000084.text	RFC822 Email Message	/File Server/File Server/HD_Loose_Documents/Folder...
☑	000680.xls	Microsoft Excel Spreadsheet	/File Server/File Server/HD_Loose_Documents/Folder...
☑	Folder003	Directory	/File Server/File Server/HD_Loose_Documents
☑	000681.jpg	JPEG/JFIF Image	/File Server/File Server/HD_Loose_Documents/Folder...
☑	000720.ppt	Microsoft PowerPoint Pres...	/File Server/File Server/HD_Loose_Documents/Folder...
☑	000682.jpg	JPEG/JFIF Image	/File Server/File Server/HD_Loose_Documents/Folder...
☑	[Unnamed Cont...	Microsoft OLE2 Package File	/File Server/File Server/HD_Loose_Documents/Folder...
☑	000683.doc	Microsoft Word Document	/File Server/File Server/HD_Loose_Documents/Folder...
☑	[Unnamed Cont...	Microsoft OLE2 Package File	/File Server/File Server/HD_Loose_Documents/Folder...
☑	000684.jpg	JPEG/JFIF Image	/File Server/File Server/HD_Loose_Documents/Folder...
☑	000721.csv	Plain Text	/File Server/File Server/HD_Loose_Documents/Folder...
☑	000685.doc	Microsoft Word Document	/File Server/File Server/HD_Loose_Documents/Folder...
☑	000748.pdf	Portable Document Format	/File Server/File Server/HD_Loose_Documents/Folder...
☑	000686.doc	Microsoft Word Document	/File Server/File Server/HD_Loose_Documents/Folder...
☑	000753.pdf	Portable Document Format	/File Server/File Server/HD_Loose_Documents/Folder...
☑	000687.doc	Microsoft Word Document	/File Server/File Server/HD_Loose_Documents/Folder...
☑	000757.pdf	Portable Document Format	/File Server/File Server/HD_Loose_Documents/Folder...

Selected 5,758 of 5,758 items.

000084.text

Path: File Server → File Server → HD_Loose_Documents → Folder001

Duplicates: Exact (0) Near (0)

Similar items: High (9) Medium (9) Low (9)

Related items: Thread (1)

Date:
From: "diane dolstad [ddolstad@hotmail.com]"
To: rule-comments@sec.gov
Cc:
Bcc:
Subject: S7-14-03

I am a stockholder who supports your new proposal, S7-14-03, on diclosure of how directors are selected. I want to know so I can invest wisely. Thank you

K Diane Dolstad
4705 Old Military Rd
Tenino WA 98589

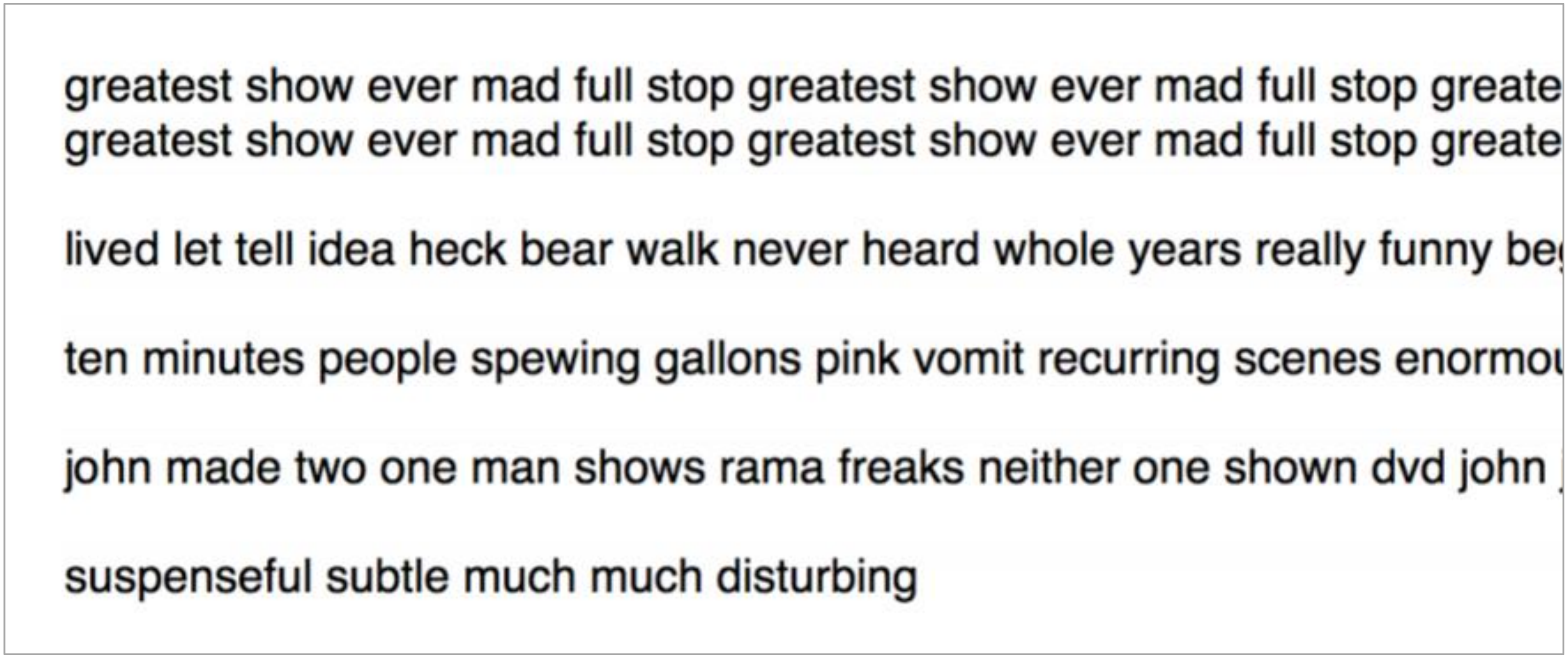
Tired of spam? Get advanced junk mail protection with MSN 8.
http://join.msn.com/?page=features/junkmail

Review and Tag: 000084.text

□ Apply same tags to all **family** items (1)
□ Apply same tags to all **duplicate** items (1)

https://www.nuix.com

Anomalies in large documents sets can be detected by unsupervised methods



greatest show ever mad full stop greatest show ever mad full stop greate
greatest show ever mad full stop greatest show ever mad full stop greate

lived let tell idea heck bear walk never heard whole years really funny be

ten minutes people spewing gallons pink vomit recurring scenes enormou

john made two one man shows rama freaks neither one shown dvd john

suspenseful subtle much much disturbing

Most anomalous reviews in the IMBD test set according to CVDD

https://www.aclweb.org/anthology/P19-1398.pdf

Thank You!

Mark Cieliebak

ciel@zhaw.ch

Image References

1. https://i.pinimg.com/474x/87/be/b8/87beb8207b44f4bb55e6757083a15794.jpg
2. https://www.vo.eu/de/wp-content/uploads/sites/2/2017/12/513691410-300x300.jpg
3. https://blog.ciat.cgiar.org/wp-content/uploads/magazines.jpg
4. https://cdn.unitycms.io/image/ocroped/2001,2000,1000,1000,0,0/QSrtW7o1c-g/CW-IGygkarqAp7vHAlvyw9.jpg
5. https://blog.ipleaders.in/wp-content/uploads/2019/04/download-2.png
6. https://www.packtpub.com/books/content/introduction-clustering-and-unsupervised-learning
7. https://neoformix.com/2009/cwc_WorldNews2008.png
8. https://carrotsearch.com/lingo4g/
9. https://www.aclweb.org/anthology/P19-1398.pdf
10. https://www.researchgate.net/figure/X-Ways-Forensics-after-the-processing-of-the-forensic-image_fig4_258332973
11. http://wallpapers-3d.ru/sstorage/53/2011/02/1100211451139523.jpg