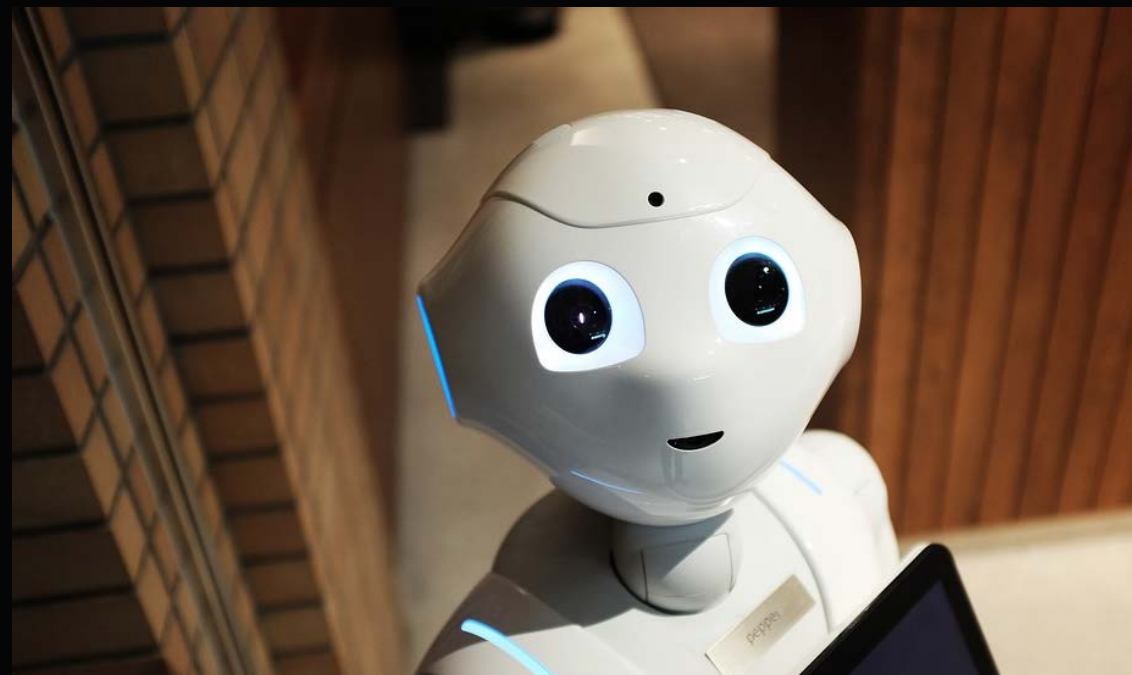


Can machines read sentences like Geoscientists do?

Paul H. Cleverley Ph.D.
Associate Lecturer
Robert Gordon University, Aberdeen, Scotland UK.

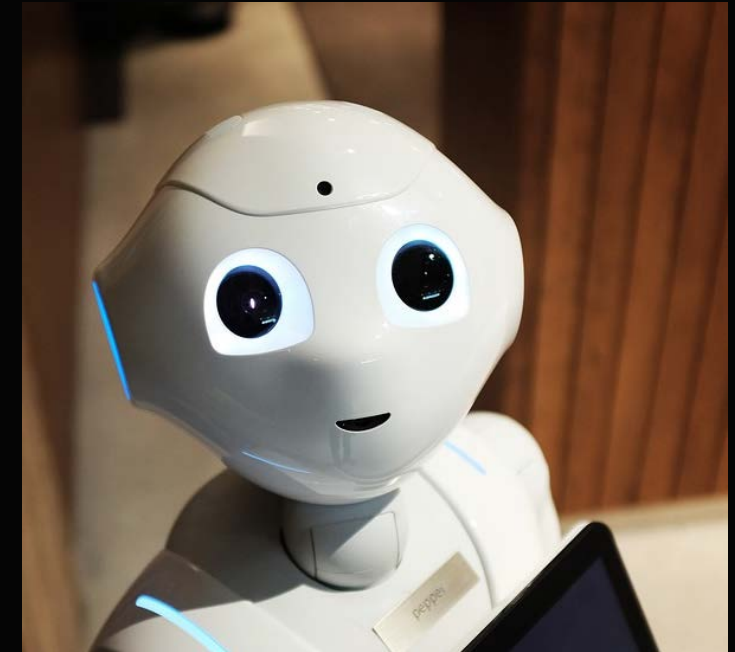


FORCE – Machine Learning on Subsurface Data Symposium 20th Sep 2018

Can machines read sentences like geoscientists do?

Assumptions – The 3 laws¹ of geoscience text analytics

- 1. Law of Non-Equivalence:** Algorithms will not read or understand sentences in the same way as *geoscientists*, because they are not geoscientists. It will be different, for better and for worse. Algorithms will have their own biases – but they may differ from our own ones which may assist and be helpful.
- 2. Law of Human Cognitive Capacity Limitations:** The amount of potentially relevant geoscience information available exceeds our cognitive capacity to read it and detect subtle patterns. It is likely that this limits the discovery of surprising, insightful and valuable connections, supporting the use case for algorithms to assist geoscientists and MEMEX devices.
- 3. Law of Inflated Expectations:** Myth and propaganda serving technological interests, may blind some of us, some of the time, to Law number #1



¹ Inspired by Isaac Asimov 's Three Law's of Robotics
Alan Turing (1950) Imitation Game
Vannevar Bush (1945)
Ellul (1954) The Technological Society

Background - Cognitive Bias

"A compelling narrative fosters an illusion of inevitability" Daniel Kahneman

"People generally see what they look for, and hear what they listen for" Harper Lee

COGNITIVE BIAS

(Rose 2016)

Premature selection of theory

Personal hubris

Lack of perspective

Lack of imagination

Laziness

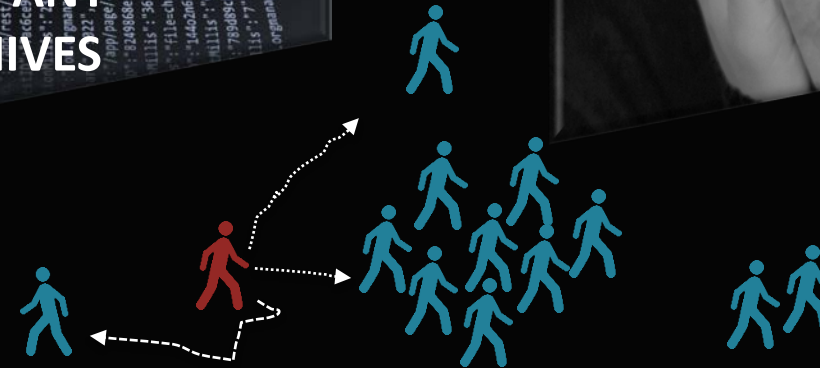
Excessive self-interest



GroupThink is to be avoided

BIG DATA
SENTIMENT
ANALYSIS FROM
EXTERNAL
LITERATURE/
COMPANY
ARCHIVES

CHALLENGE
SURPRISE US?



.... but being aware of independently stacked opinion in literature may challenge our own biases



**ROBERT GORDON
UNIVERSITY ABERDEEN**

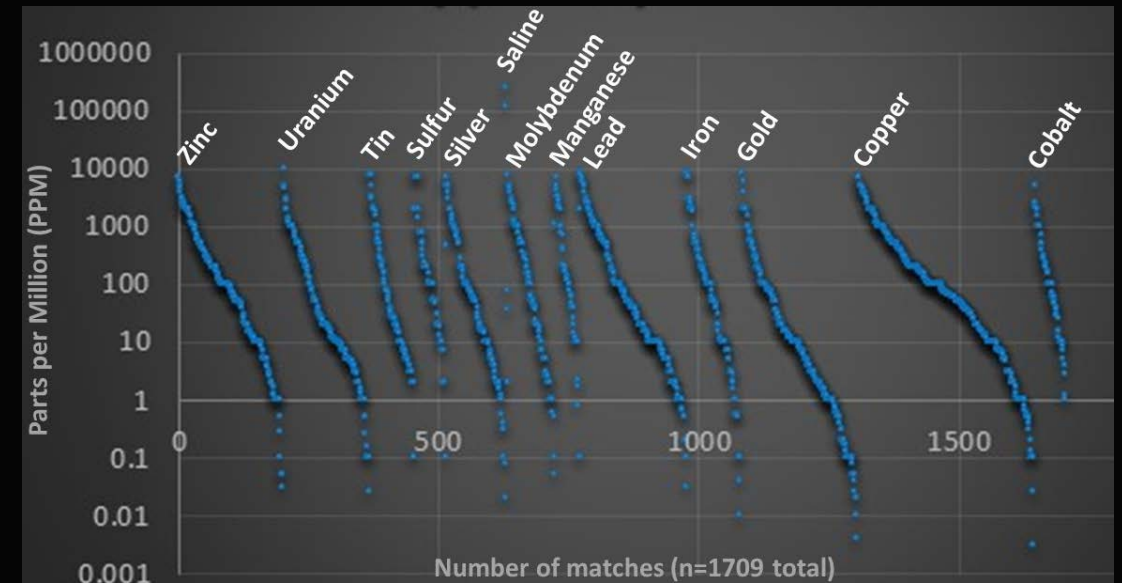
Text and Data Mining – Rule Based

Spatializing entities/concepts and associations
e.g. 'mentions' of Pre-Cambrian

C shows histograms of δC analyses of anthraxolite and organic carbon in Precambrian sedimentary rocks, graphite in noncalcareous Precambrian schists and gneisses, and vein graphites from Ceylon, Montana, and New Hampshire. From these data, it is clear that

'Extracting integer and float data from unstructured text
e.g. ppm is an association with a chemical element

Monroe, where a ridge on the east meets the plain of Sevier River. It is discharging about 30 gallons minute of water at temperatures range from per that 135 ϕ to 146 ϕ F; an analysis indicates that it contains 0.41 ppm of manganese.



Cleverley (2017) Data courtesy of the Society of Economic Geology via GeoscienceWorld

Unsupervised machine learning: Word Vectors

Question: Which is the most similar play to the 'Green Formation'?

	Word count of co-occurrences		
Play	Reservoir	Source Rock	Trap
Green Formation			
Beano Formation			
Nemo Formation			
...			

Answer: Nemo is the most similar play based on latent patterns in text

	Word count of co-occurrences			
Play	shoreface	well rounded	fractured	...
Green Formation: Reservoir	5	3	2	
Beano Formation: Reservoir	0	1	2	
Nemo Formation: Reservoir	1	4	3	
...				

$$\frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

Green Formation v Beano Formation = $(5*0+3*1+0*5+2*2)/((\text{SQRT}(5*5+3*3+0*0+2*2))x(\text{SQRT}(0*0+1*1+5*5+2*2))) = 0.2075$

Green Formation v Nemo Formation = $(5*1+3*4+0*0+2*3)/((\text{SQRT}(5*5+3*3+0*0+2*2))x(\text{SQRT}(1*1+4*4+0*0+3*3))) = 0.7333$

Unsupervised machine learning: analogues

I need an analogue for..

Analogue	Overall	Reservoir	Source	Trap
Play 47	<u>82%</u>	<u>91%</u>	<u>87%</u>	<u>63%</u>
Play 12	<u>80%</u>	<u>79%</u>	<u>82%</u>	<u>45%</u>
Play 45a	<u>65%</u>	<u>72%</u>	<u>15%</u>	<u>75%</u>
Play 3	<u>61%</u>	<u>91%</u>	<u>15%</u>	<u>25%</u>
Play 102	<u>58%</u>	<u>47%</u>	<u>20%</u>	<u>33%</u>
Play 56	<u>54%</u>	<u>34%</u>	<u>39%</u>	<u>62%</u>
Play 32	<u>51%</u>	<u>81%</u>	<u>18%</u>	<u>15%</u>
...				

Co-occurrence Algorithms



Text Corpus (Documents)

To date I have only tested concepts with geoscientists on Lithostratigraphic Units not the whole play concept. Promising results:

"I input the xxx Formation that I studied in Tunisia and it returned a lateral equivalent (in Libya) that I had not come across before."

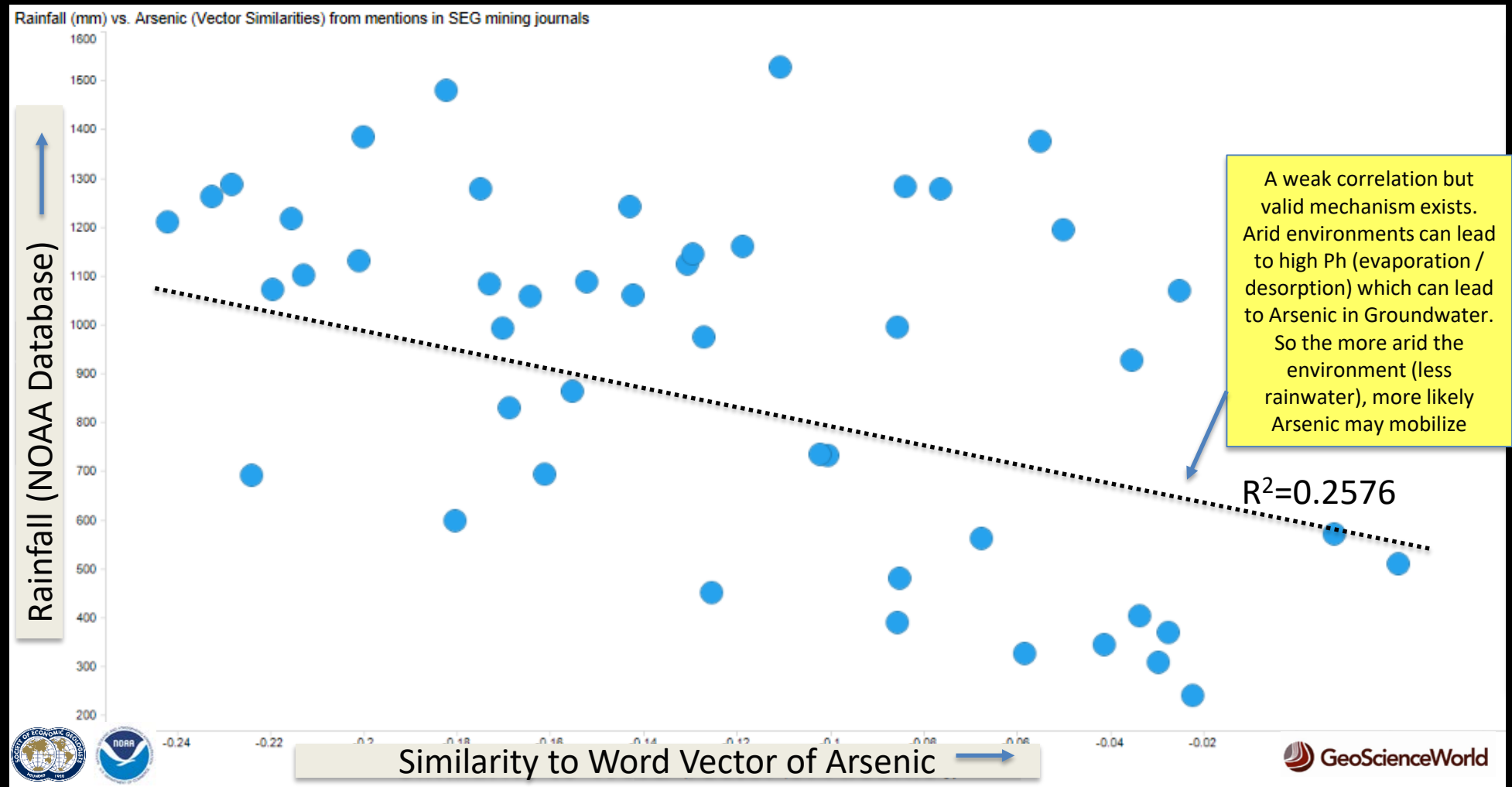
Geologist, Multi-National Oil and Gas Company

Hypothesis Tester (Economic Geology Example)

US States ●

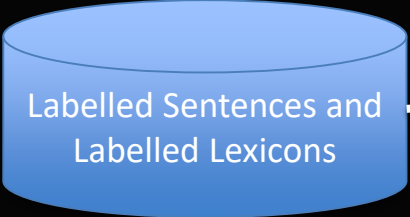
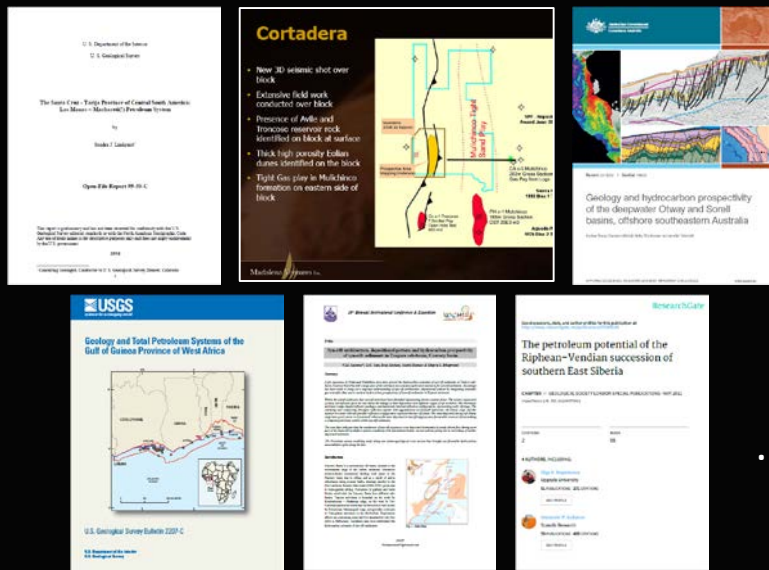
Y-Axis = annual rainfall data from NOAA

X-Axis = similarity of word vectors of all US States in SEG text (6,000 papers) to the word vector of Arsenic in SEG text



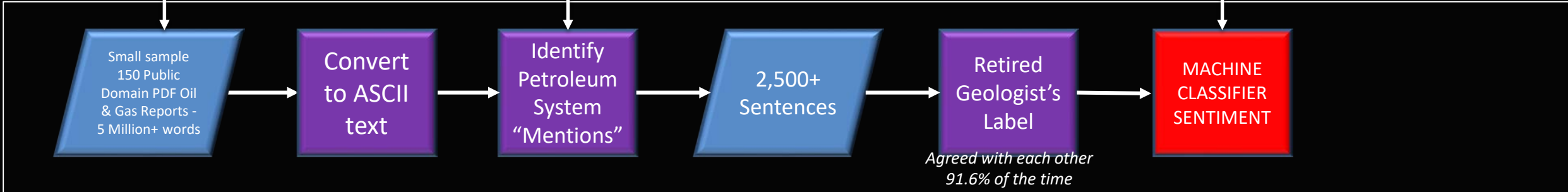
Cleverley, P.H. (2017) Text Analytics meets Geoscience. New Mexico, April 2017

Supervised Machine Learning: Sentiment Analysis

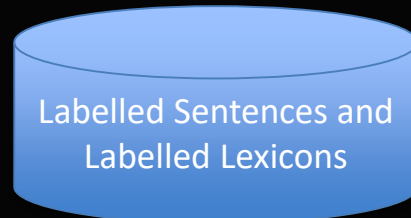
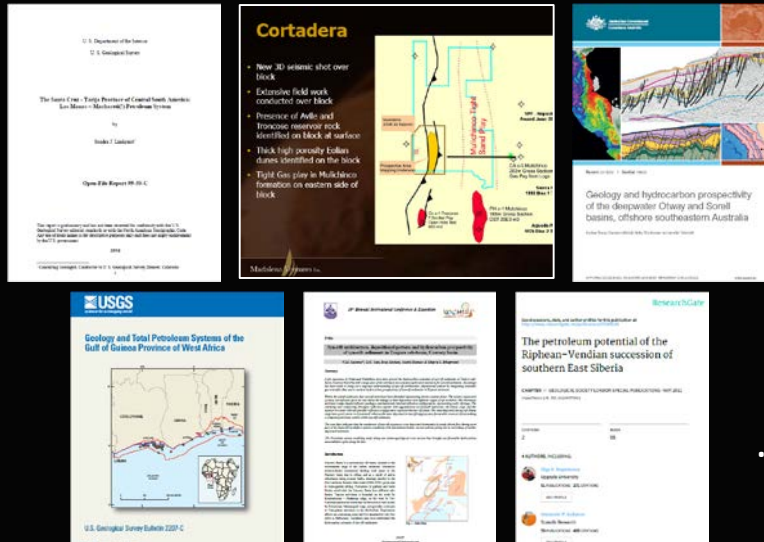


```
1 #####  
2 # Geoscience Aware Sentiment Analyser (GAZER)  
3 # Python 2.7 64BIT  
4 # Paul H Cleverley February 2018  
5 # Machine Learning Custom Feature-Set/Lexicon Training Bayesian Model  
6 # Knowledge engineering: with skipgrams for word order, Numerical Rules  
7 # Natural Language Processing (NLP) Part of Speech recognition, Plurals, Negation  
8 #####  
9 import nltk  
10 from nltk import word_tokenize  
11 import re  
12 import functools  
13 from nltk.tokenize import word_tokenize  
14 from nltk.util import skipgrams  
15 import functools  
16 import re  
17 from os import chdir  
18 chdir('C:/Users/pppp/Desktop/paulhjan')  
19 # nltk supervised sentiment classifier  
20 from nltk.classifiers import NaiveBayesClassifier  
21 with open('FINAL_GEO-LEXICON-SHIP_Feb18.csv', 'r') as fp:  
22     cl = NaiveBayesClassifier(fp, format='csv')  
23 print('done!')  
24 cutoff=0.54  
25 t=3  
26 from nltk import TextBlob  
27 def fopen(filename):  
28     '''Opens a text file, reads lines(), returns content'''  
29     fobj = open(filename, 'r')  
30     data = fobj.readlines()  
31     fobj.close()  
32     return data  
33 NEUTFP_POS=0  
34 NEUTFP_NEG=0  
35 POSFP_NEUT=0  
36 POSFP_POS=0  
37 NEGFP_POS=0  
38 NEGFP_NEG=0  
39 #####  
40 # Main Program  
41 # Get list of files to process  
42 # Loop over files  
43 # Get text from file  
44 # Tokenize text  
45 # Create skipgrams  
46 # Classify sentiment  
47 # Print results  
48 # Update counters  
49 # Print summary  
50 #####  
51 # Summary  
52 print('Total POSITIVE test items: ', '250', 'FP POS: ', '25', 'accuracy: ', '0.9')  
53 print('Total NEGATIVE test items: ', '250', 'FP NEG: ', '16', 'accuracy: ', '0.936')  
54 print('Accuracy 2 categories: ', '0.9426666666666667')  
55 print('Accuracy 3 categories: ', '0.9426666666666667')  
56 #####  
57 # Accuracy  
58 print('NEUTOT: ', '250', 'success: ', '180', 'FP POS: ', '54', 'NEG: ', '16')  
59 print('NEUTOT: ', '250', 'success: ', '227', 'FP POS: ', '11', 'NEUTRAL: ', '12')  
60 print('POSTOT: ', '250', 'success: ', '225', 'FP NEG: ', '6', 'NEUTRAL: ', '19')  
61 print('ACCRAC3: ', '0.9426666666666667')
```

Python Code – Geoscience Aware Sentiment Analyzer (GAZER)



Supervised Machine Learning: Sentiment Analysis



```

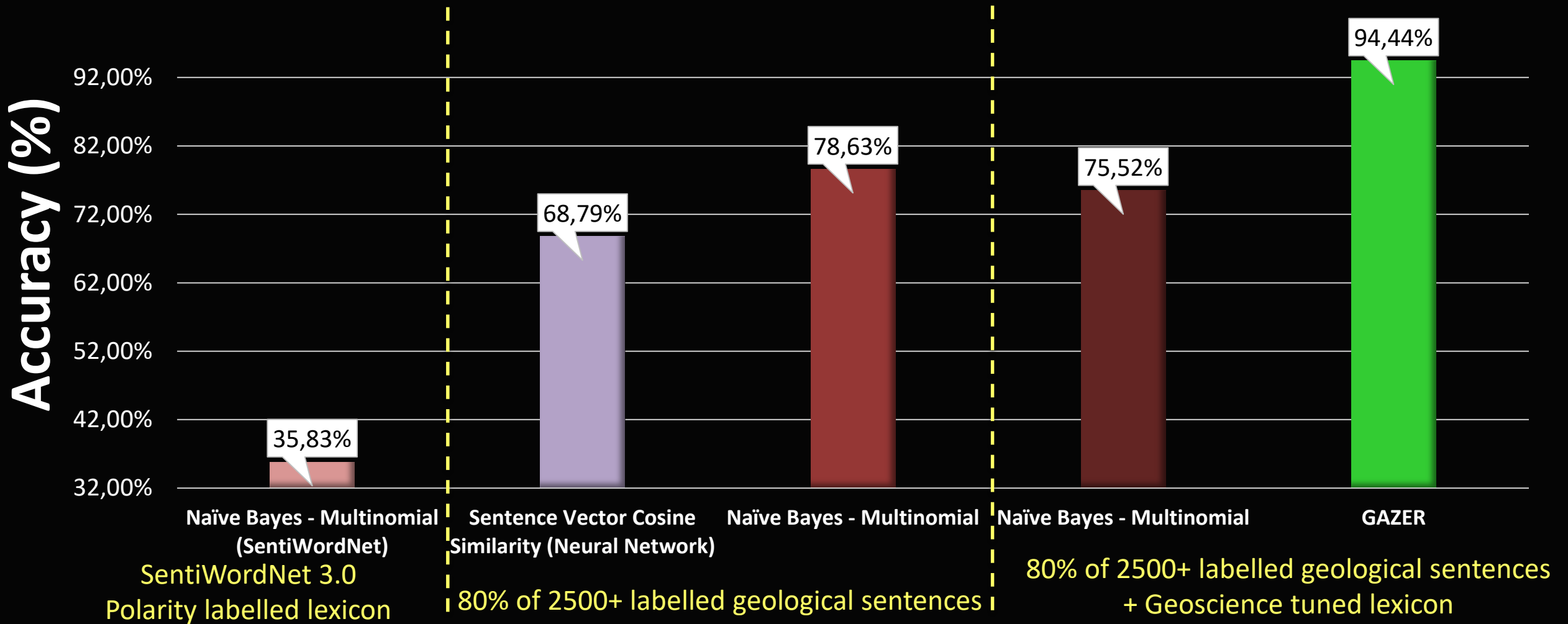
Spyder (Python 2.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
C:\Users\pepp\Desktop\pauljan
C:\Users\pepp\Desktop\pauljan\government_2018_dip_V2_Neutral_POS_NLP2-19Feb.py
File Edit Search Source Run Debug Consoles Projects Tools View Help
C:\Users\pepp\Desktop\pauljan\government_2018_dip_V2_Neutral_POS_NLP2-19Feb.py
1 #####
2 # Government Analyst (GATTA) V1.0
3 # Python 2.7 64BIT
4 # Paul H Cleverley February 2018
5 # Machine Learning Custom Feature-Set/Lexicon Training Bayesian Model
6 # Knowledge engineering: with skipgrams for word order, numerical rules
7 # Natural Language Processing (NLP) Part of Speech recognition, Plurals, Negation
8 #####
9 import nltk
10 from nltk import word_tokenize
11 import re
12 import functools
13 from nltk.tokenize import word_tokenize
14 from nltk.util import skipgrams
15 import functools
16 import re
17 from os import chdir
18 chdir('C:\Users\pepp\Desktop\pauljan')
19 # nltk.classifiers sentiment classifier
20 from nltk.classifiers import NaiveBayesClassifier
21 with open('FINAL_GEO-LEXICON-SHIP_Feb18.csv', 'r') as fp:
22     cl = NaiveBayesClassifier(fp, format="csv")
23 print ("done")
24 cutoff=0.54
25 t=3
26 from textblob import TextBlob
27 def fopen(filename):
28     """Opens a text file, readlines(), returns content"""
29     fobj = open(filename, "r")
30     data = fobj.readlines()
31     fobj.close()
32     return data
33 NEUTFP_POS=0
34 NEUTFP_NEG=0
35 POSFP_NEG=0
36 POSFP_NEUT=0
37 NEGFP_POS=0
38 #####
39 #####
40 #####
41 #####
42 #####
43 #####
44 #####
45 #####
46 #####
47 #####
48 #####
49 #####
50 #####
51 #####
52 #####
53 #####
54 #####
55 #####
56 #####
57 #####
58 #####
59 #####
60 #####
61 #####
62 #####
63 #####
64 #####
65 #####
66 #####
67 #####
68 #####
69 #####
70 #####
71 #####
72 #####
73 #####
74 #####
75 #####
76 #####
77 #####
78 #####
79 #####
80 #####
81 #####
82 #####
83 #####
84 #####
85 #####
86 #####
87 #####
88 #####
89 #####
90 #####
91 #####
92 #####
93 #####
94 #####
95 #####
96 #####
97 #####
98 #####
99 #####
100 #####
101 #####
102 #####
103 #####
104 #####
105 #####
106 #####
107 #####
108 #####
109 #####
110 #####
111 #####
112 #####
113 #####
114 #####
115 #####
116 #####
117 #####
118 #####
119 #####
120 #####
121 #####
122 #####
123 #####
124 #####
125 #####
126 #####
127 #####
128 #####
129 #####
130 #####
131 #####
132 #####
133 #####
134 #####
135 #####
136 #####
137 #####
138 #####
139 #####
140 #####
141 #####
142 #####
143 #####
144 #####
145 #####
146 #####
147 #####
148 #####
149 #####
150 #####
151 #####
152 #####
153 #####
154 #####
155 #####
156 #####
157 #####
158 #####
159 #####
160 #####
161 #####
162 #####
163 #####
164 #####
165 #####
166 #####
167 #####
168 #####
169 #####
170 #####
171 #####
172 #####
173 #####
174 #####
175 #####
176 #####
177 #####
178 #####
179 #####
180 #####
181 #####
182 #####
183 #####
184 #####
185 #####
186 #####
187 #####
188 #####
189 #####
190 #####
191 #####
192 #####
193 #####
194 #####
195 #####
196 #####
197 #####
198 #####
199 #####
200 #####
201 #####
202 #####
203 #####
204 #####
205 #####
206 #####
207 #####
208 #####
209 #####
210 #####
211 #####
212 #####
213 #####
214 #####
215 #####
216 #####
217 #####
218 #####
219 #####
220 #####
221 #####
222 #####
223 #####
224 #####
225 #####
226 #####
227 #####
228 #####
229 #####
230 #####
231 #####
232 #####
233 #####
234 #####
235 #####
236 #####
237 #####
238 #####
239 #####
240 #####
241 #####
242 #####
243 #####
244 #####
245 #####
246 #####
247 #####
248 #####
249 #####
250 #####
251 #####
252 #####
253 #####
254 #####
255 #####
256 #####
257 #####
258 #####
259 #####
260 #####
261 #####
262 #####
263 #####
264 #####
265 #####
266 #####
267 #####
268 #####
269 #####
270 #####
271 #####
272 #####
273 #####
274 #####
275 #####
276 #####
277 #####
278 #####
279 #####
280 #####
281 #####
282 #####
283 #####
284 #####
285 #####
286 #####
287 #####
288 #####
289 #####
290 #####
291 #####
292 #####
293 #####
294 #####
295 #####
296 #####
297 #####
298 #####
299 #####
300 #####
301 #####
302 #####
303 #####
304 #####
305 #####
306 #####
307 #####
308 #####
309 #####
310 #####
311 #####
312 #####
313 #####
314 #####
315 #####
316 #####
317 #####
318 #####
319 #####
320 #####
321 #####
322 #####
323 #####
324 #####
325 #####
326 #####
327 #####
328 #####
329 #####
330 #####
331 #####
332 #####
333 #####
334 #####
335 #####
336 #####
337 #####
338 #####
339 #####
340 #####
341 #####
342 #####
343 #####
344 #####
345 #####
346 #####
347 #####
348 #####
349 #####
350 #####
351 #####
352 #####
353 #####
354 #####
355 #####
356 #####
357 #####
358 #####
359 #####
360 #####
361 #####
362 #####
363 #####
364 #####
365 #####
366 #####
367 #####
368 #####
369 #####
370 #####
371 #####
372 #####
373 #####
374 #####
375 #####
376 #####
377 #####
378 #####
379 #####
380 #####
381 #####
382 #####
383 #####
384 #####
385 #####
386 #####
387 #####
388 #####
389 #####
390 #####
391 #####
392 #####
393 #####
394 #####
395 #####
396 #####
397 #####
398 #####
399 #####
400 #####
401 #####
402 #####
403 #####
404 #####
405 #####
406 #####
407 #####
408 #####
409 #####
410 #####
411 #####
412 #####
413 #####
414 #####
415 #####
416 #####
417 #####
418 #####
419 #####
420 #####
421 #####
422 #####
423 #####
424 #####
425 #####
426 #####
427 #####
428 #####
429 #####
430 #####
431 #####
432 #####
433 #####
434 #####
435 #####
436 #####
437 #####
438 #####
439 #####
440 #####
441 #####
442 #####
443 #####
444 #####
445 #####
446 #####
447 #####
448 #####
449 #####
450 #####
451 #####
452 #####
453 #####
454 #####
455 #####
456 #####
457 #####
458 #####
459 #####
460 #####
461 #####
462 #####
463 #####
464 #####
465 #####
466 #####
467 #####
468 #####
469 #####
470 #####
471 #####
472 #####
473 #####
474 #####
475 #####
476 #####
477 #####
478 #####
479 #####
480 #####
481 #####
482 #####
483 #####
484 #####
485 #####
486 #####
487 #####
488 #####
489 #####
490 #####
491 #####
492 #####
493 #####
494 #####
495 #####
496 #####
497 #####
498 #####
499 #####
500 #####
501 #####
502 #####
503 #####
504 #####
505 #####
506 #####
507 #####
508 #####
509 #####
510 #####
511 #####
512 #####
513 #####
514 #####
515 #####
516 #####
517 #####
518 #####
519 #####
520 #####
521 #####
522 #####
523 #####
524 #####
525 #####
526 #####
527 #####
528 #####
529 #####
530 #####
531 #####
532 #####
533 #####
534 #####
535 #####
536 #####
537 #####
538 #####
539 #####
540 #####
541 #####
542 #####
543 #####
544 #####
545 #####
546 #####
547 #####
548 #####
549 #####
550 #####
551 #####
552 #####
553 #####
554 #####
555 #####
556 #####
557 #####
558 #####
559 #####
560 #####
561 #####
562 #####
563 #####
564 #####
565 #####
566 #####
567 #####
568 #####
569 #####
570 #####
571 #####
572 #####
573 #####
574 #####
575 #####
576 #####
577 #####
578 #####
579 #####
580 #####
581 #####
582 #####
583 #####
584 #####
585 #####
586 #####
587 #####
588 #####
589 #####
590 #####
591 #####
592 #####
593 #####
594 #####
595 #####
596 #####
597 #####
598 #####
599 #####
600 #####
601 #####
602 #####
603 #####
604 #####
605 #####
606 #####
607 #####
608 #####
609 #####
610 #####
611 #####
612 #####
613 #####
614 #####
615 #####
616 #####
617 #####
618 #####
619 #####
620 #####
621 #####
622 #####
623 #####
624 #####
625 #####
626 #####
627 #####
628 #####
629 #####
630 #####
631 #####
632 #####
633 #####
634 #####
635 #####
636 #####
637 #####
638 #####
639 #####
640 #####
641 #####
642 #####
643 #####
644 #####
645 #####
646 #####
647 #####
648 #####
649 #####
650 #####
651 #####
652 #####
653 #####
654 #####
655 #####
656 #####
657 #####
658 #####
659 #####
660 #####
661 #####
662 #####
663 #####
664 #####
665 #####
666 #####
667 #####
668 #####
669 #####
670 #####
671 #####
672 #####
673 #####
674 #####
675 #####
676 #####
677 #####
678 #####
679 #####
680 #####
681 #####
682 #####
683 #####
684 #####
685 #####
686 #####
687 #####
688 #####
689 #####
690 #####
691 #####
692 #####
693 #####
694 #####
695 #####
696 #####
697 #####
698 #####
699 #####
700 #####
701 #####
702 #####
703 #####
704 #####
705 #####
706 #####
707 #####
708 #####
709 #####
710 #####
711 #####
712 #####
713 #####
714 #####
715 #####
716 #####
717 #####
718 #####
719 #####
720 #####
721 #####
722 #####
723 #####
724 #####
725 #####
726 #####
727 #####
728 #####
729 #####
730 #####
731 #####
732 #####
733 #####
734 #####
735 #####
736 #####
737 #####
738 #####
739 #####
740 #####
741 #####
742 #####
743 #####
744 #####
745 #####
746 #####
747 #####
748 #####
749 #####
750 #####
751 #####
752 #####
753 #####
754 #####
755 #####
756 #####
757 #####
758 #####
759 #####
760 #####
761 #####
762 #####
763 #####
764 #####
765 #####
766 #####
767 #####
768 #####
769 #####
770 #####
771 #####
772 #####
773 #####
774 #####
775 #####
776 #####
777 #####
778 #####
779 #####
780 #####
781 #####
782 #####
783 #####
784 #####
785 #####
786 #####
787 #####
788 #####
789 #####
790 #####
791 #####
792 #####
793 #####
794 #####
795 #####
796 #####
797 #####
798 #####
799 #####
800 #####
801 #####
802 #####
803 #####
804 #####
805 #####
806 #####
807 #####
808 #####
809 #####
810 #####
811 #####
812 #####
813 #####
814 #####
815 #####
816 #####
817 #####
818 #####
819 #####
820 #####
821 #####
822 #####
823 #####
824 #####
825 #####
826 #####
827 #####
828 #####
829 #####
830 #####
831 #####
832 #####
833 #####
834 #####
835 #####
836 #####
837 #####
838 #####
839 #####
840 #####
841 #####
842 #####
843 #####
844 #####
845 #####
846 #####
847 #####
848 #####
849 #####
850 #####
851 #####
852 #####
853 #####
854 #####
855 #####
856 #####
857 #####
858 #####
859 #####
860 #####
861 #####
862 #####
863 #####
864 #####
865 #####
866 #####
867 #####
868 #####
869 #####
870 #####
871 #####
872 #####
873 #####
874 #####
875 #####
876 #####
877 #####
878 #####
879 #####
880 #####
881 #####
882 #####
883 #####
884 #####
885 #####
886 #####
887 #####
888 #####
889 #####
890 #####
891 #####
892 #####
893 #####
894 #####
895 #####
896 #####
897 #####
898 #####
899 #####
900 #####
901 #####
902 #####
903 #####
904 #####
905 #####
906 #####
907 #####
908 #####
909 #####
910 #####
911 #####
912 #####
913 #####
914 #####
915 #####
916 #####
917 #####
918 #####
919 #####
920 #####
921 #####
922 #####
923 #####
924 #####
925 #####
926 #####
927 #####
928 #####
929 #####
930 #####
931 #####
932 #####
933 #####
934 #####
935 #####
936 #####
937 #####
938 #####
939 #####
940 #####
941 #####
942 #####
943 #####
944 #####
945 #####
946 #####
947 #####
948 #####
949 #####
950 #####
951 #####
952 #####
953 #####
954 #####
955 #####
956 #####
957 #####
958 #####
959 #####
960 #####
961 #####
962 #####
963 #####
964 #####
965 #####
966 #####
967 #####
968 #####
969 #####
970 #####
971 #####
972 #####
973 #####
974 #####
975 #####
976 #####
977 #####
978 #####
979 #####
980 #####
981 #####
982 #####
983 #####
984 #####
985 #####
986 #####
987 #####
988 #####
989 #####
990 #####
991 #####
992 #####
993 #####
994 #####
995 #####
996 #####
997 #####
998 #####
999 #####
1000 #####
    
```

TEXT SENTENCE AUTOMATICALLY EXTRACTED FROM DOCUMENTS

POLARITY

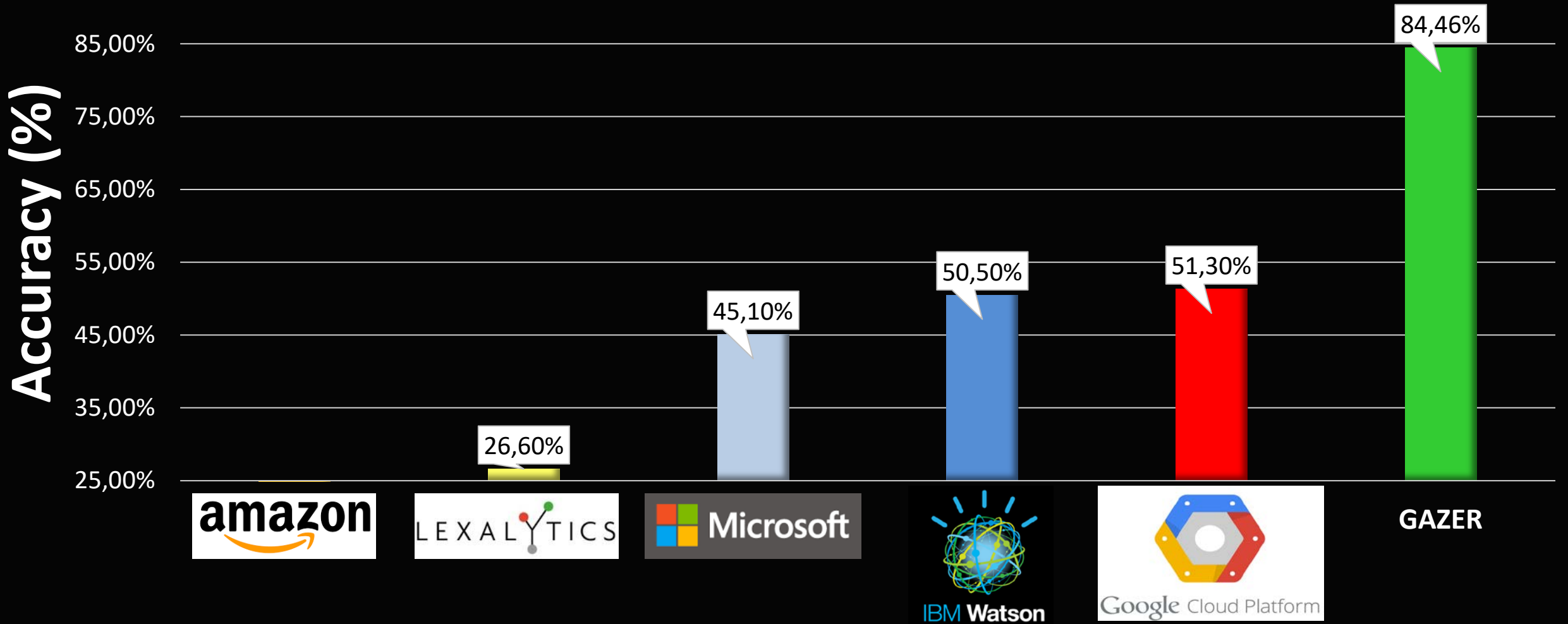
characterization of fluvial and aeolian reservoirs	Neutral
source rocks are thought to have expelled oil after trap formation	Positive
reservoir sands are not a problem in this area	Positive
..shows several attractive fault related traps but sedimentological and stratigraphic studies indicate that there would be poor reservoir characteristics..	Negative
the ro values between 0.5 and 0.7% indicate low source-rock grade	Negative
consequently, a key subsequent step in exploration was to establish the presence of source rocks in the basin	Neutral

Results – Accuracy Comparison 2 Categories POS v NEG



- Comparing too State-of-the-art in the literature for generic sentiment analysis: Sentence Vector (Le & Mikolov 2014) gave 92.6% with 25,000 Movie Reviews as a training set

Results – Using 750 Test Set 3 Cat. (POS-NEG-NEUT)



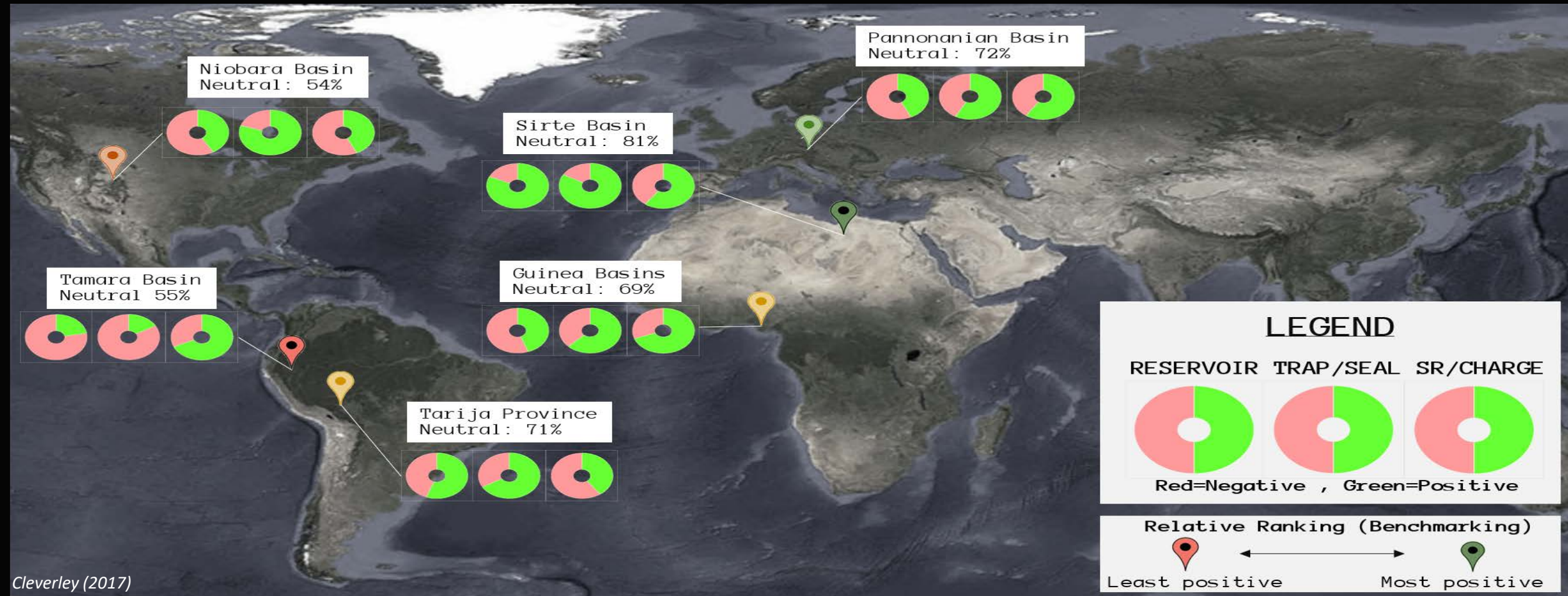
Published data sets

750 labelled sentences (POS, NEG, NEUT)

for test benchmarking classifier performance on Github:

<https://github.com/phcleverley/Geoscience-Sentiment-Research>

Data driven insights



Cleverley (2017)

"...a really nice way to capture literature without reading it. It would be nice to run this method on all the AAPG publications of a few separate basins and see if the graphs reflect our basic understanding of these basins. This could become a very powerful method in understanding and visualizing the current state of knowledge." Exploration Geologist (December 2017)

Summary and future work

- **Extracting text & numerical data from documents is only half the story..**
- **Algorithms can approach ‘geoscientist-like’ accuracy on narrow classification tasks like sentiment analysis, but can process exponentially greater volumes of info.**
- **Finding analogues is a ‘similarity’ challenge. Identifying what similarity dimensions and characteristics appeal to geoscientists (and in which contexts and why) is an area for further research.**



Cleverley & Burnett (2014) Study of 53 geoscientists using various stimulants on touchscreens