Sentiment Analysis of Social Media Posts

Niels Dommerholt

JDriven

- Small (~ 40 employees) consulting company in Nieuwegein, The Netherlands
- Specialized in the JVM stack: Java, Groovy, Scala, etc.
- A sparring partner for the top 250 organizations in The Netherlands

Agenda

- Reddit dataset
- Apache Spark
- Sentiment Analysis
- Results

Reddit

- Huge "Community of Communities"
- Founded in 2005
- Anyone can create and participate in communities (called "subreddits")
- Driven by user submitted content (links, funny pictures, in-depth discussions)
- Mostly user-moderated

Data set

- Full dataset: ~ 1.7 billion user contributions between 2007 and June 2015
- Well over 1TB in size
- On Google BigQuery (links on http://bit.ly/1HmZ9No)

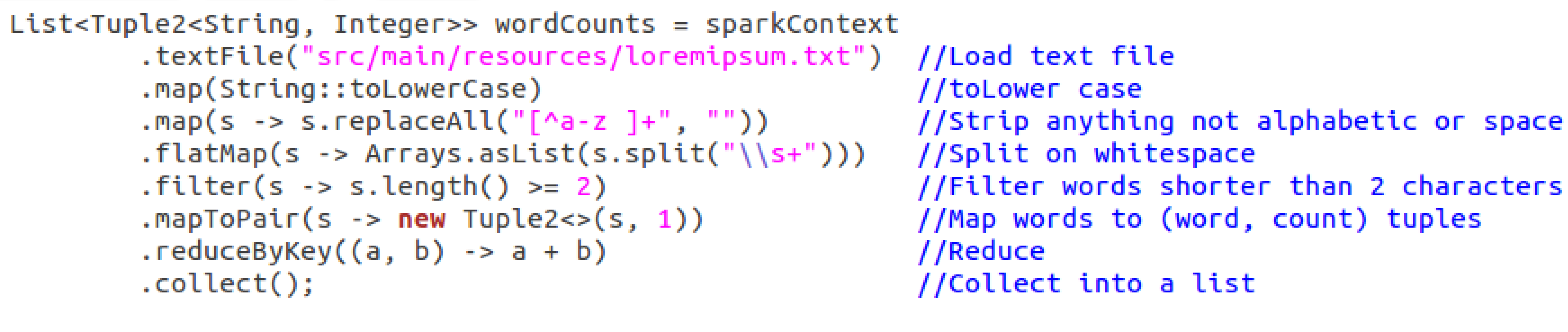
- Extract I used: user contributions in January 2015
- 53.8 million comments in one month (on average 20 per second!)
- 5.4 GB compressed JSON (31GB uncompressed)

Apache Spark

- Open Source Big Data Processing Framework

- Fast
- Smart
- Easy to use
- Generic

Hello Spark World



```java
List<Tuple2<String, Integer>> wordCounts = sparkContext
        .textFile("src/main/resources/loremipsum.txt")  //Load text file
        .map(String::toLowerCase)                        //toLower case
        .map(s -> s.replaceAll("[^a-z ]+", ""))          //Strip anything not alphabetic or space
        .flatMap(s -> Arrays.asList(s.split("\\s+")))    //Split on whitespace
        .filter(s -> s.length() >= 2)                    //Filter words shorter than 2 characters
        .mapToPair(s -> new Tuple2<>(s, 1))              //Map words to (word, count) tuples
        .reduceByKey((a, b) -> a + b)                    //Reduce
        .collect();                                      //Collect into a list
```

Sentiment Analysis 101

“Sentiment Analysis is the process of detecting the contextual polarity of text.”

- In short: trying to find out if a piece of text is neutral, positive of negative
- Typically involves annotating large volumes of text with 'polarity' scores

Example:

- “awesome” score: + 2.0
- “horrible” score: -2.5
- “not bad at all” score: +1.5

My Approach

- Using an existing free word list: http://sentiwordnet.isti.cnr.it
- There are better ones, but they are often expensive!
- Analyse comments word for word adding up the “scores”
- The “sentiment” of a comment is it's normalized score (total score / total number of words)

Example

- Input: “Such an abhorrent sense of betrayal”

- “Such (-0.125) an (-0.125) abhorrent (-0.75) sense (-0.125) of betrayal (-0.25)”
  - Score: -1.375
  - Normalized: Score / Words = -1.375 / 6 = ~ −0,23

Most popular subreddits

#	Sub	Comments
1	askreddit	4,712,795
2	nfl	932,460
3	funny	930,098
4	leagueoflegends	904,297
5	pics	778,942
6	worldnews	670,872
7	todayilearned	599,295
8	destinythegame	587,774
9	adviceanimals	577,463
10	videos	570,938

- Total 53.8 million comments
- AskReddit by far the most popular (9%)
- Around ~ 47,000 subs with at least one comment in Jan 2015

Most productive authors

#	Author	Comments
1	automoderator	233144
2	politicbot	61889
3	autowikibot	22599
4	tweetposter	16325
5	havoc_bot	14186
6	doctor-kitten	13830
7	mtgcardfetcher	12306
8	imgurtranscriber	10302
9	rpbot	10014
10	marvelvsdc00	9090

- Most of the most productive authors are bots

Most used words

the	in	not	can	would
i	s	was	like	there
to	for	are	your	one
a	t	if	he	don
and	this	they	at	about
it	on	my	what	get
you	but	as	me	we
of	have	just	m	from
that	be	or	all	out
is	with	so	do	an

- “The” has 60 million occurences
- “An” has 4.6 million
- Useful way to find “stop words”

Positive comments per sub

#	Sub	Comments
1	askreddit	282876
2	funny	58697
3	pics	53860
4	nfl	53844
5	random_acts_of_amazon	53726
6	leagueoflegends	52312
7	nba	33717
8	gonewild	33522
9	pcmasterrace	32689

- Number of comments with score > 0.1
- Results not normalized

Negative comments per sub

#	Sub	Comments
1	askreddit	275244
2	funny	56409
3	nfl	50370
4	pics	43044
5	leagueoflegends	39701
6	adviceanimals	28628
7	videos	27195
8	nba	26553
9	wtf	26473
10	todayilearned	26143

- Number of comments with score < -0.1
- Results not normalized

Comment sentiment per day

	Total	**Positive**		**Negative**	
MONDAY	6573954	405741	6.2%	285709	4.3%
TUESDAY	6881857	451814	**6.6%**	295575	4.3%
WEDNESDAY	6965360	425338	6.1%	294937	4.2%
THURSDAY	7986816	494134	6.2%	346129	4.3%
FRIDAY	8402932	519293	6.2%	365291	4.3%
SATURDAY	7097303	457755	6.4%	321889	4.5%
SUNDAY	6025960	393887	6.5%	280998	**4.7%**

Uses of Sentiment Analysis

- Detect positive / negative customer reviews
- Detect positive / negative customer opinions
- How are customers responding to changes we made?
- Discarding subjective information
- “Flame” detection
- Identify bias in news sources

Challenges in Sentiment Analysis

- Understanding context in sentences is hard
- Annotating large bodies of text to train a Natural Language processing system is a lot of work
- Not an exact science: “A study from the University of Pittsburgh shows that humans can only agree on whether or not a sentence has the correct sentiment, 80% of the time.”

Conclusion

- I hope you enjoyed the show!
- Links to the data and information as well as the sample code can be found in the repo:

  https://github.com/nielsutrecht/jfall-sentiment/

  (or scan the QR code)
- For my contact details:

  http://niels.nu/

THANK
YOU