

Phone Sentiment Analysis with Web Crawl Data

Tuomo Kareoja

Alert! Analytics

October 30, 2019

Agenda

Summary of Findings

- People Like iPhone More

- ...but the Training Data for Samsung Galaxy is Sketchy

- ...and the Websites Mentioning iPhone are Weird

What We Did?

- Count Instances of Word and Word Combinations from Websites

- Train Models on Manually Labeled Data and Apply These to the Crawled Data

What to Do Better Next Time?

- Make Sure That the Labeling Makes Sense

- Aggressive Limitations on What Sites to Crawl

Conclusions

People Like iPhone More

Only keeping sites with at least two mentions of the phone

...but the Training Data for Samsung Galaxy is Sketchy

Picture of sentiment and mentions of the phone

...and the Websites Mentioning iPhone are Weird

Wordcloud

Count Instances of Word and Word Combinations from Websites

Common Crawl Mentions of phones Positive and negative overall mentions and specific to certain attributes

Picture of model comparison

Train Models on Manually Labeled Data and Apply These to the Crawled Data

Make Sure That the Labeling Makes Sense

Was this really done by hand and if so, what were the principals of evaluation?

If done programmatically, where is the code?

How is it possible that sites that don't contain any mention of the the device can have emotional sentiment towards it?

Aggressive Limitations on What Sites to Crawl

What sites are we trying to find?

Forums are tricky but might be valuable

News are problematic because of clickbait

Conclusions

1. iPhone has a more positive sentiment, but this might be because of (fraudulent) giveaways and such
2. ...but if people would want the Galaxy phones as much as iPhone we would expect there to be a similar number of these sites for Samsung Galaxy phones and this does not seem to be the case
3. More carefully documented labeling needed in the future
4. Web crawling should be even more targeted

The End

Questions?