Analyzing Unstructured Big Text Data

Sertac Karaman
Assistant Professor of Aeronautics and Astronautics
Laboratory for Information and Decision Systems
Institute for Data, Systems, and Society
Massachusetts Institute of Technology

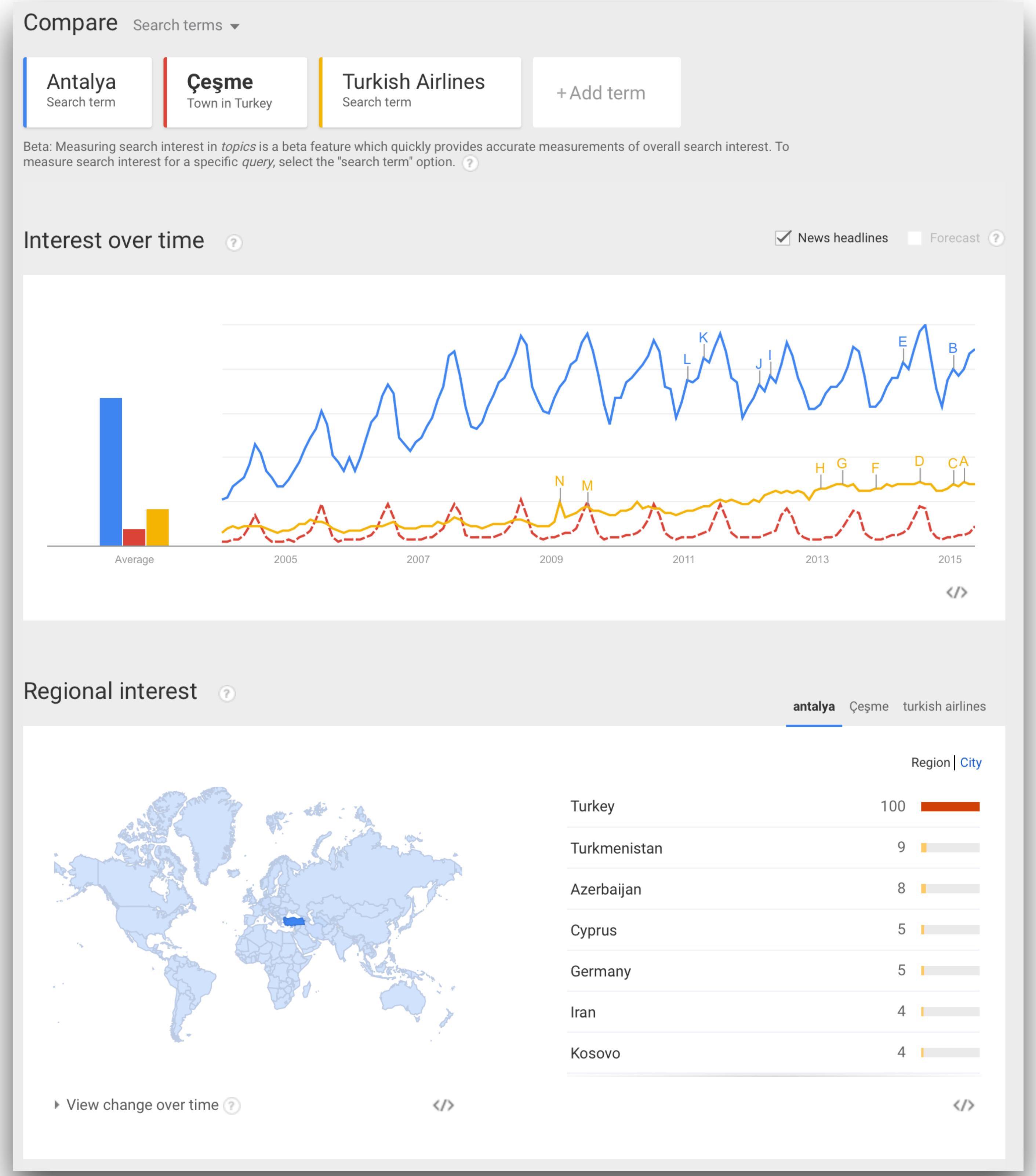
Why Analyze Text Data

- Very little data is in the traditional “relational” format.
- Most of the World’s data (roughly 80%) is unstructured, same holds for many large scale businesses:
    - Emails, social media posts, surveys, requests, notes, comments, call centers, news stories, research papers, legal memoranda, websites, …
- Sometimes information with substantial business value can only be found inside unstructured text.
- Businesses that master text analytics may be able to get ahead quickly.

Why is Text Analytics Hard?

- Language is ambiguous
  - Context is required to clarify
  - The same words may mean different things (homographs)
    - Bear (verb) - to support, carry
    - Bear (noun) - a large animal
  - Different words can mean the same things (synonyms)
- Language is subtle (e.g., sarcasm)
- Concept/Word extraction usually results huge “number of dimensions”
  - Thousands of new fields
  - Each field typically has low information content
- Misspellings, abbreviations, spelling variants, …, particularly in twitter
  - Customer vs. cstmr

The Basic Stuff: Count Analysis



Compare Search terms ▾

Antalya Search term	Çeşme Town in Turkey	Turkish Airlines Search term	+ Add term

Beta: Measuring search interest in *topics* is a beta feature which quickly provides accurate measurements of overall search interest. To measure search interest for a specific *query*, select the "search term" option. ?

Interest over time ?

☑ News headlines □ Forecast ?

Regional interest ?

antalya Çeşme turkish airlines

Region | City

Turkey	100
Turkmenistan	9
Azerbaijan	8
Cyprus	5
Germany	5
Iran	4
Kosovo	4

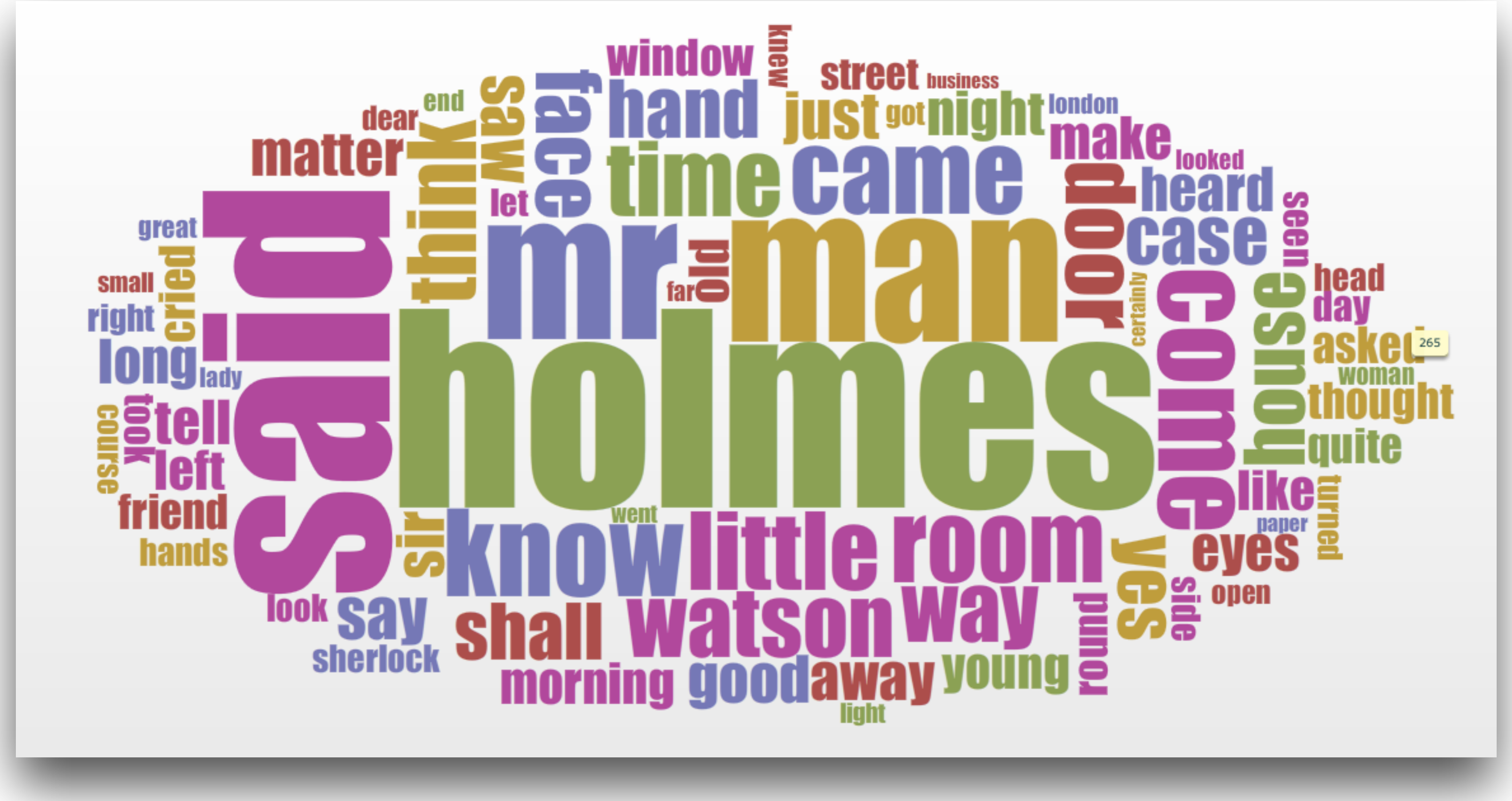
▸ View change over time ?

Trends in Google Search

http://www.google.com/trends

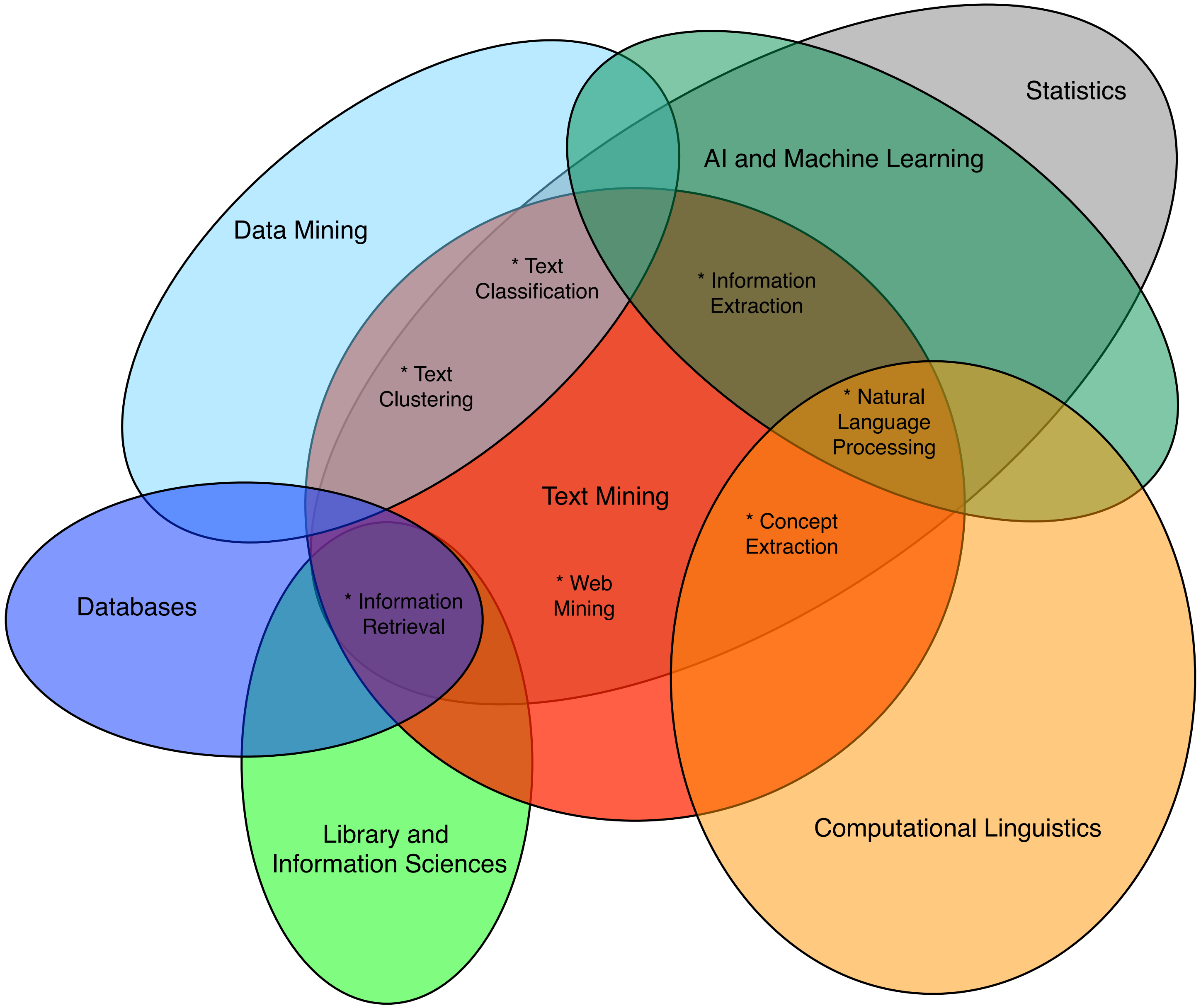
The Basic Stuff: Frequency Analysis

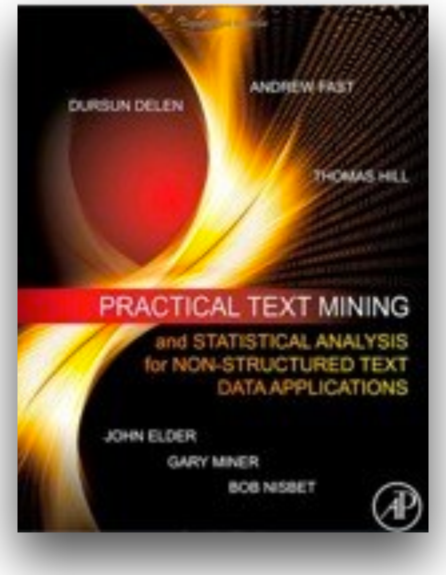
- Cloud plots are utilized for frequency analysis



Frequency of the most frequent words in Sherlock Holmes

Text Analytics (or Text Mining) is considered the “next frontier” by many in the research community





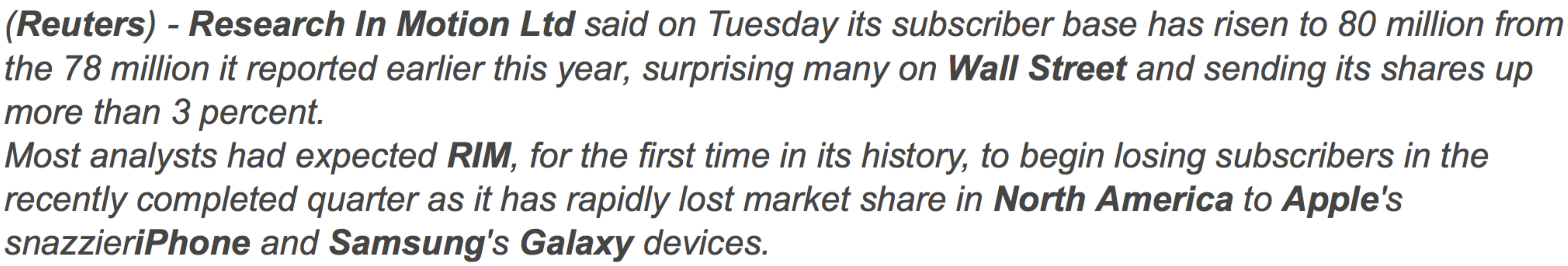
The field is interdisciplinary, and it is vast!

Some of the Main Categories

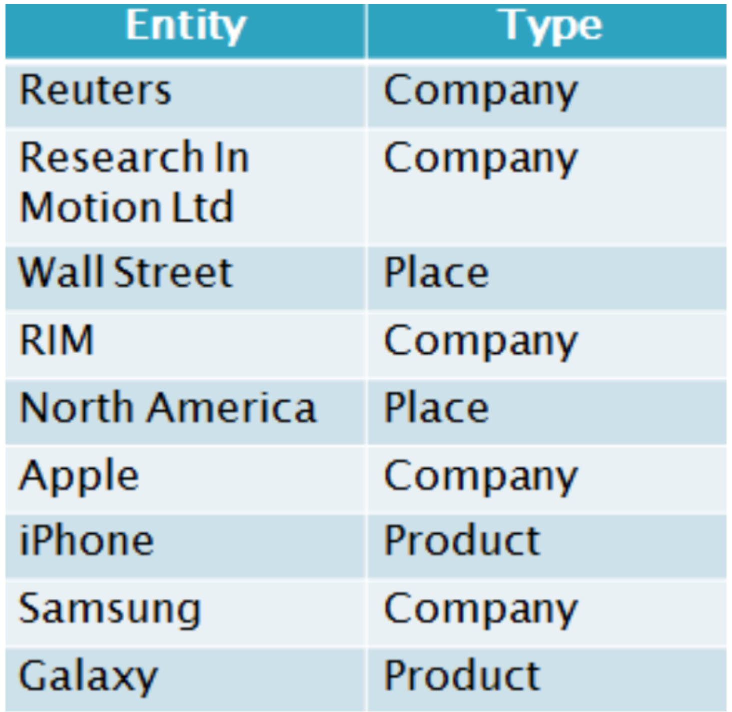
- What is possible with text analytics today?
  - Named entities extraction
  - Document summarization
  - Theme extraction
  - Concept extraction
  - Sentiment analysis

Named Entities Extraction

- **Named Entities Extraction** helps answer the question "who, what and where" is being discussed.



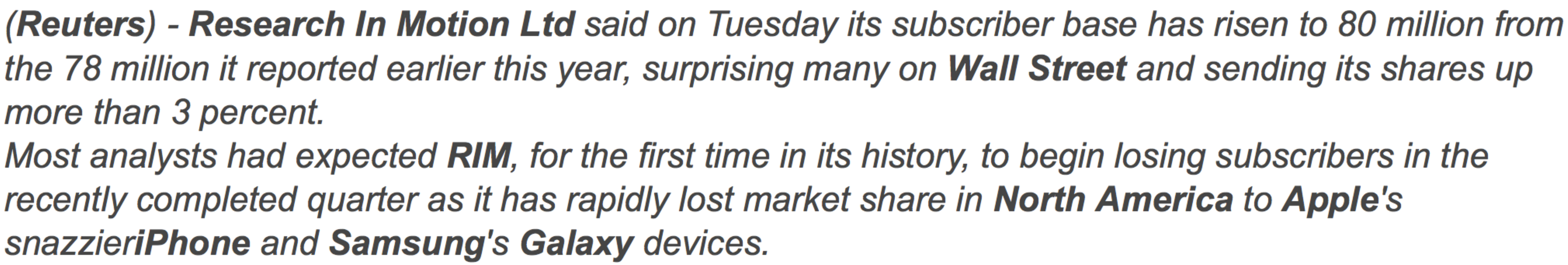
*(**Reuters**) - **Research In Motion Ltd** said on Tuesday its subscriber base has risen to 80 million from the 78 million it reported earlier this year, surprising many on **Wall Street** and sending its shares up more than 3 percent.*
*Most analysts had expected **RIM**, for the first time in its history, to begin losing subscribers in the recently completed quarter as it has rapidly lost market share in **North America** to **Apple**'s snazzieri**Phone** and **Samsung**'s **Galaxy** devices.*



Entity	Type
Reuters	Company
Research In Motion Ltd	Company
Wall Street	Place
RIM	Company
North America	Place
Apple	Company
iPhone	Product
Samsung	Company
Galaxy	Product

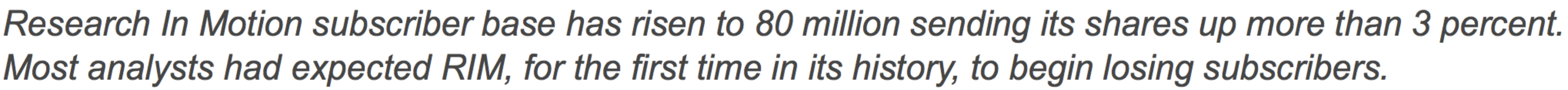
Document Summarization

- **Document summarization** is the creation of a shortened version of a text by a computer program.



*(**Reuters**) - **Research In Motion Ltd** said on Tuesday its subscriber base has risen to 80 million from the 78 million it reported earlier this year, surprising many on **Wall Street** and sending its shares up more than 3 percent.*
*Most analysts had expected **RIM**, for the first time in its history, to begin losing subscribers in the recently completed quarter as it has rapidly lost market share in **North America** to **Apple**'s snazzieri**Phone** and **Samsung**'s **Galaxy** devices.*

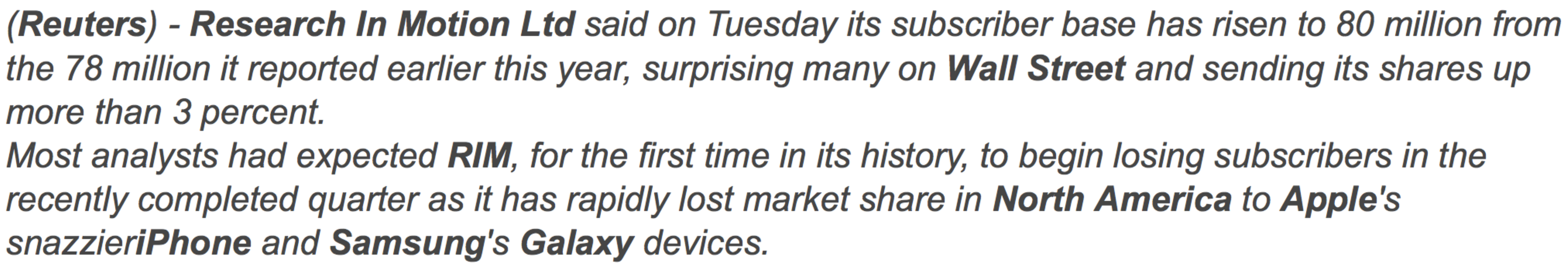
Summarized by software:



Research In Motion subscriber base has risen to 80 million sending its shares up more than 3 percent.
Most analysts had expected RIM, for the first time in its history, to begin losing subscribers.

Theme Extraction

- **Theme Extraction** answers the question "what are the important words being used"?



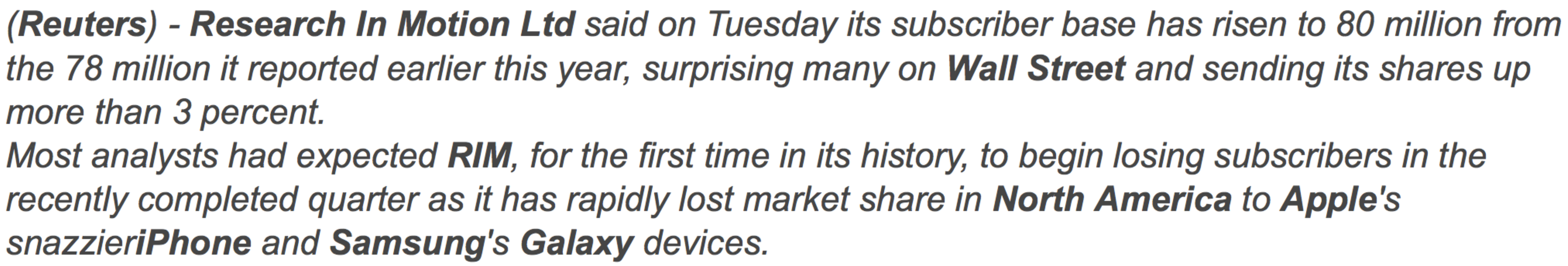
*(**Reuters**) - **Research In Motion Ltd** said on Tuesday its subscriber base has risen to 80 million from the 78 million it reported earlier this year, surprising many on **Wall Street** and sending its shares up more than 3 percent.*
*Most analysts had expected **RIM**, for the first time in its history, to begin losing subscribers in the recently completed quarter as it has rapidly lost market share in **North America** to **Apple**'s snazzieri**Phone** and **Samsung**'s **Galaxy** devices.*

The software finds:

- Subscriber base
- Shares
- Market share

Concept Extraction

- **Concept extraction** or **concept mining** is an activity that results in the extraction of concepts from artifacts. Concept Extraction answers the question: "what" are the important high level concepts?



*(**Reuters**) - **Research In Motion Ltd** said on Tuesday its subscriber base has risen to 80 million from the 78 million it reported earlier this year, surprising many on **Wall Street** and sending its shares up more than 3 percent.*
*Most analysts had expected **RIM**, for the first time in its history, to begin losing subscribers in the recently completed quarter as it has rapidly lost market share in **North America** to **Apple**'s snazzieri**Phone** and **Samsung**'s **Galaxy** devices.*

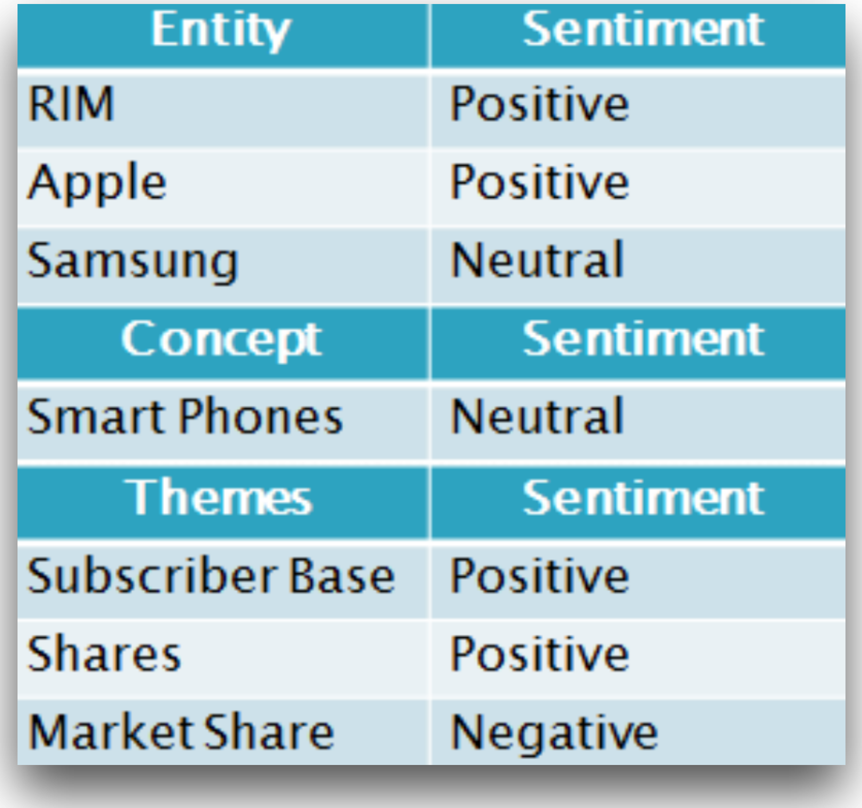
The software finds:

- Subscriber base
- Shares
- Market share

Sentiment Analysis

- **Concept extraction** or **concept mining** is an activity that results in the extraction of concepts from artifacts. Concept Extraction answers the question: "what" are the important high level concepts?

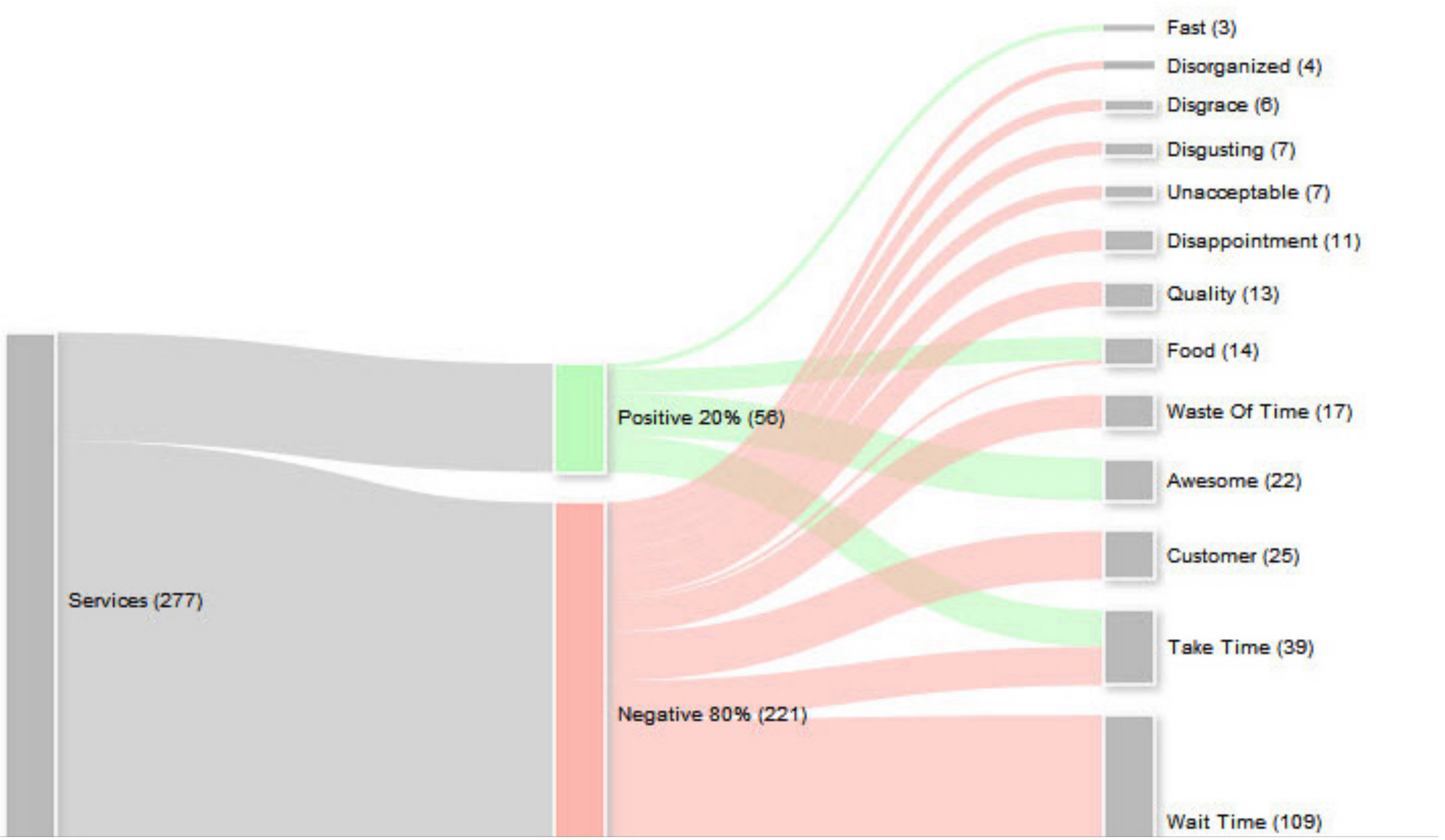
*(**Reuters**) - **Research In Motion Ltd** said on Tuesday its subscriber base has risen to 80 million from the 78 million it reported earlier this year, surprising many on **Wall Street** and sending its shares up more than 3 percent.*
*Most analysts had expected **RIM**, for the first time in its history, to begin losing subscribers in the recently completed quarter as it has rapidly lost market share in **North America** to **Apple**'s snazzieri**Phone** and **Samsung**'s **Galaxy** devices.*



Entity	Sentiment
RIM	Positive
Apple	Positive
Samsung	Neutral
Concept	**Sentiment**
Smart Phones	Neutral
Themes	**Sentiment**
Subscriber Base	Positive
Shares	Positive
Market Share	Negative

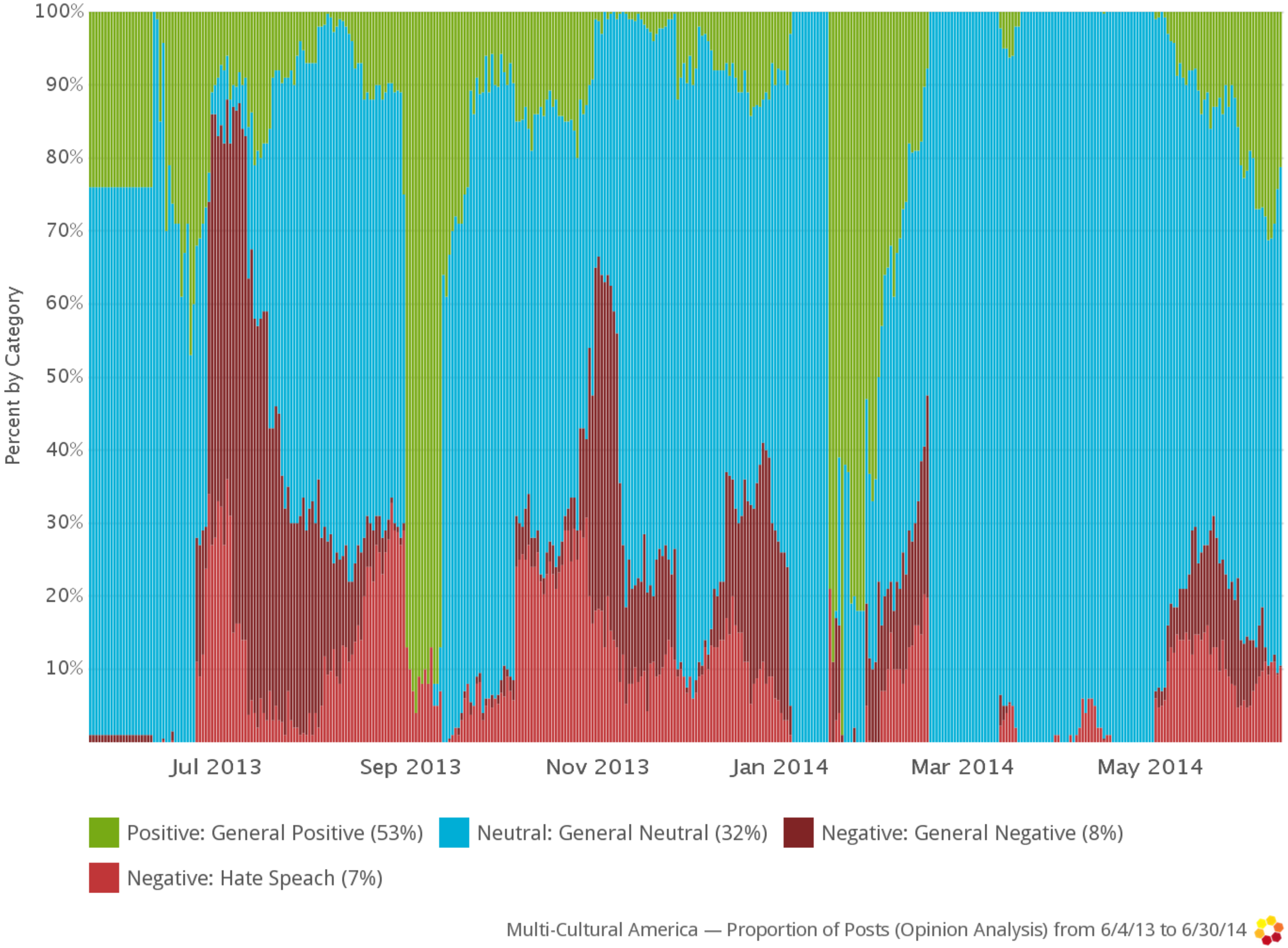
Software finds the sentiment towards all entities, concepts, and themes

Sentiment Analysis



Tell us: How did you like our new service?

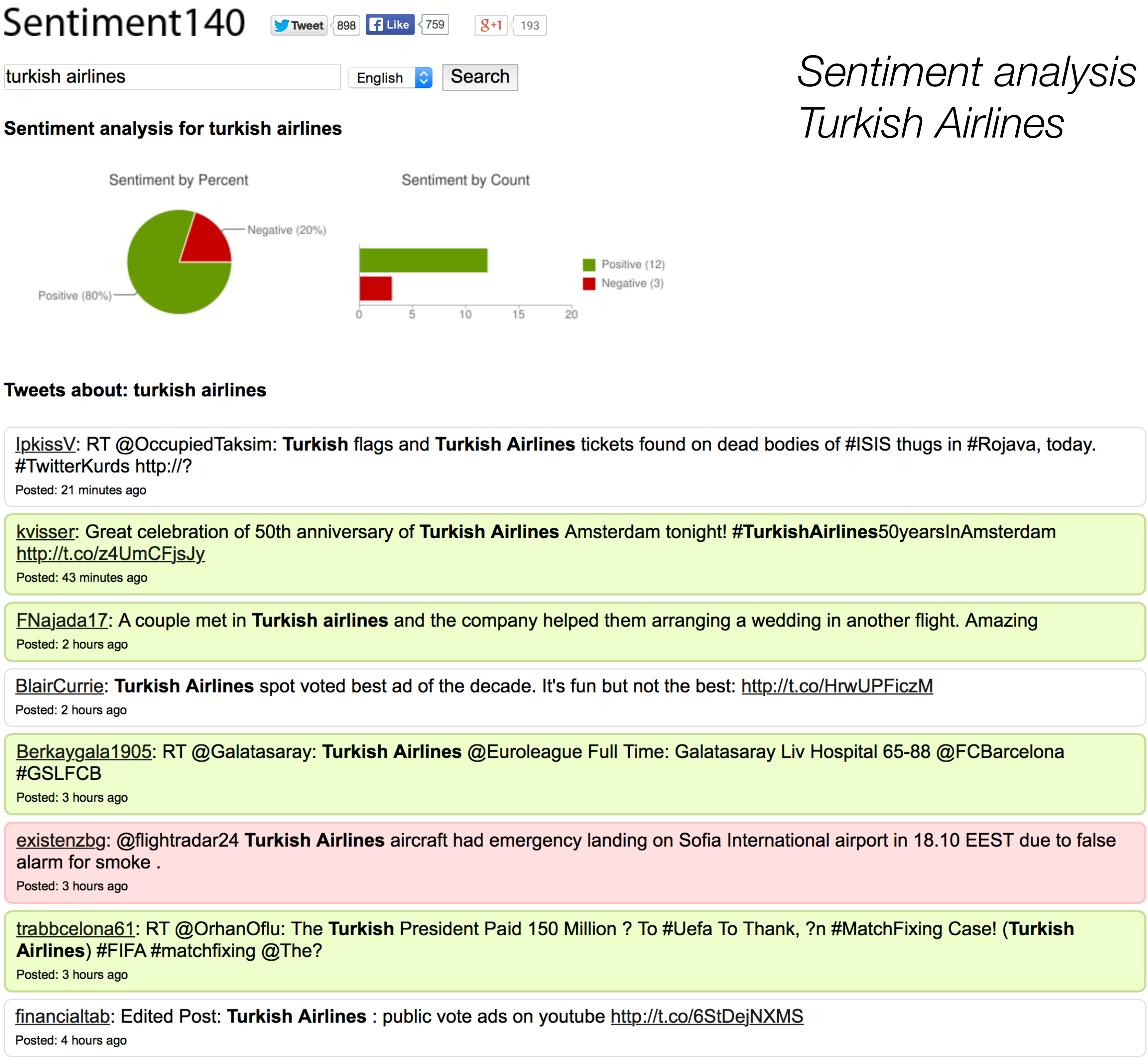
Sentiment Analysis



Should America be multi-cultural?

MIT AEROASTRO

Sentiment Analysis



turkish airlines | English | Search

Sentiment analysis for turkish airlines

Tweets about: turkish airlines

IpkissV: RT @OccupiedTaksim: **Turkish** flags and **Turkish Airlines** tickets found on dead bodies of #ISIS thugs in #Rojava, today. #TwitterKurds http://?
Posted: 21 minutes ago

kvisser: Great celebration of 50th anniversary of **Turkish Airlines** Amsterdam tonight! #**TurkishAirlines**50yearsInAmsterdam http://t.co/z4UmCFjsJy
Posted: 43 minutes ago

FNajada17: A couple met in **Turkish airlines** and the company helped them arranging a wedding in another flight. Amazing
Posted: 2 hours ago

BlairCurrie: **Turkish Airlines** spot voted best ad of the decade. It's fun but not the best: http://t.co/HrwUPFiczM
Posted: 2 hours ago

Berkaygala1905: RT @Galatasaray: **Turkish Airlines** @Euroleague Full Time: Galatasaray Liv Hospital 65-88 @FCBarcelona #GSLFCB
Posted: 3 hours ago

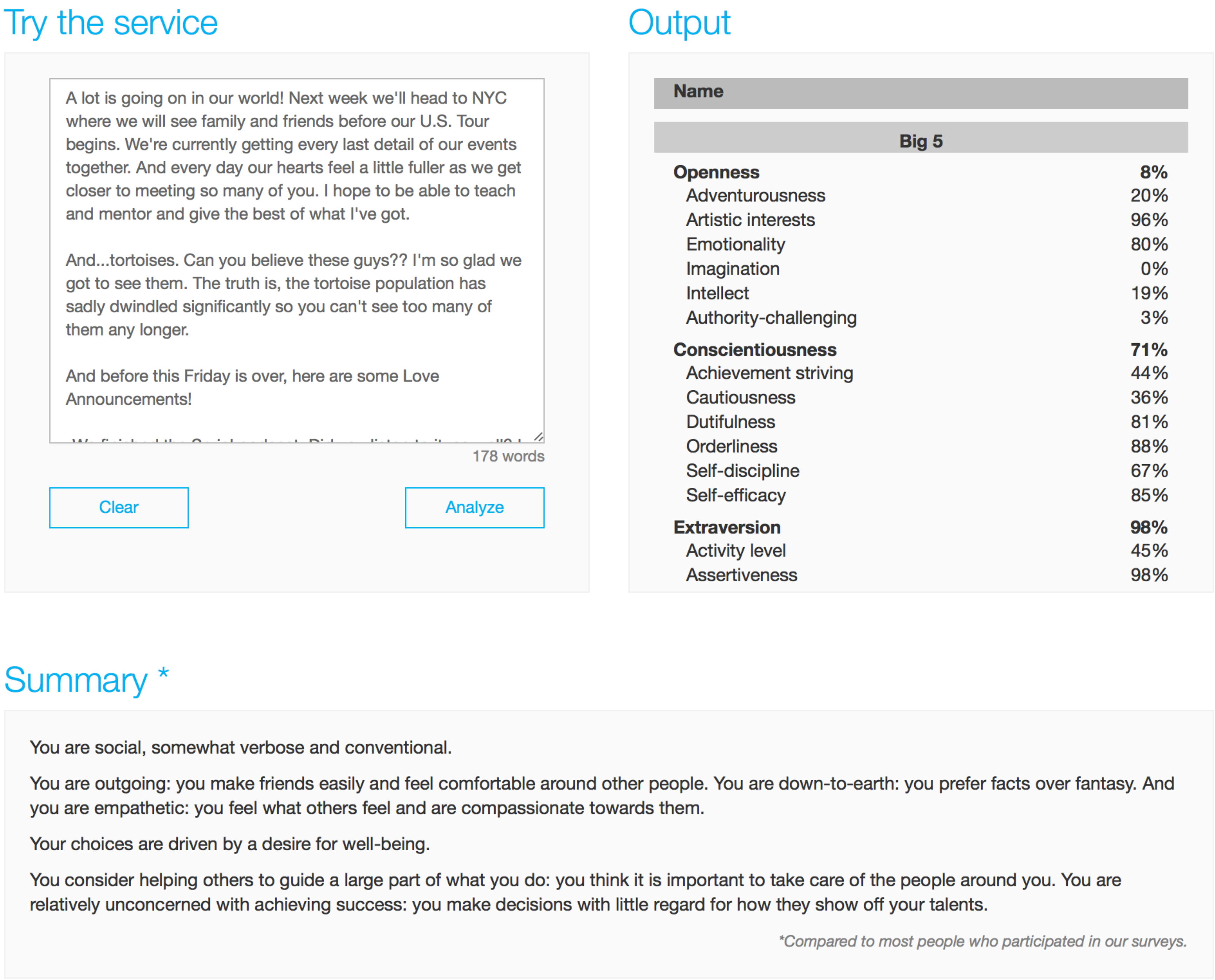
existenzbg: @flightradar24 **Turkish Airlines** aircraft had emergency landing on Sofia International airport in 18.10 EEST due to false alarm for smoke .
Posted: 3 hours ago

trabbcelona61: RT @OrhanOflu: The **Turkish** President Paid 150 Million ? To #Uefa To Thank, ?n #MatchFixing Case! (**Turkish Airlines**) #FIFA #matchfixing @The?
Posted: 3 hours ago

financialtab: Edited Post: **Turkish Airlines** : public vote ads on youtube http://t.co/6StDejNXMS
Posted: 4 hours ago

Sentiment analysis for Turkish Airlines

More Advanced Stuff:
Full “Personality Analysis” with IBM’s Watson



Try the service

A lot is going on in our world! Next week we'll head to NYC where we will see family and friends before our U.S. Tour begins. We're currently getting every last detail of our events together. And every day our hearts feel a little fuller as we get closer to meeting so many of you. I hope to be able to teach and mentor and give the best of what I've got.

And...tortoises. Can you believe these guys?? I'm so glad we got to see them. The truth is, the tortoise population has sadly dwindled significantly so you can't see too many of them any longer.

And before this Friday is over, here are some Love Announcements!

178 words

Output

Name	
Big 5	
Openness	**8%**
Adventurousness	20%
Artistic interests	96%
Emotionality	80%
Imagination	0%
Intellect	19%
Authority-challenging	3%
Conscientiousness	**71%**
Achievement striving	44%
Cautiousness	36%
Dutifulness	81%
Orderliness	88%
Self-discipline	67%
Self-efficacy	85%
Extraversion	**98%**
Activity level	45%
Assertiveness	98%

Summary *

You are social, somewhat verbose and conventional.

You are outgoing: you make friends easily and feel comfortable around other people. You are down-to-earth: you prefer facts over fantasy. And you are empathetic: you feel what others feel and are compassionate towards them.

Your choices are driven by a desire for well-being.

You consider helping others to guide a large part of what you do: you think it is important to take care of the people around you. You are relatively unconcerned with achieving success: you make decisions with little regard for how they show off your talents.

**Compared to most people who participated in our surveys.*

Watson is able to understand personality from text!

Try: *http://watson-um-demo.mybluemix.net*

What can we do beyond sentiment analysis?

- **Sentiment analysis** over twitter has become one of the important applications, as it became extremely hard for businesses to track the response that they collect from millions of their customers.

- Recent advances (in the last couple of years) in machine learning enabled unprecedented advances in text analytics.