IBM Social Media Analytics – Text Analysis on Big Data
Agenda

Social Media Analytics: Scope and Myths
What to measure in Social?
Analysis approaches and Challenges
Our Text Analysis Environment
Three areas of Social Media Analytics

- **Content**
  - Tweets, forum posts, blogs, video comments,...
  - Files shared in Collaboration Suites

- **People**
  - Geographic, Demographic, Behavioral Profiles
  - Expertise and Influence

- **Relationships**
  - How does content spread?
  - How do people interact?
Some myths about Social Media Analytics

- It’s all about twitter and facebook
  - Buying decisions are often made based on reviews, blog entries, forum discussions

- It’s all about detecting the next outrage
  - Detecting consumer sentiment is just one of many insights
  - Social media is a good way to learn what people like about products / services

- Predict elections, sales demand,...by looking at social media alone
  - Social media can be a valuable addition, but never a replacement for planning, surveys,...
  - Social analysis results need to be integrated with internal data for more relevance

- You need to analyze petabytes in nanoseconds for relevant results
  - It’s the analysis depth that counts – some analyses yield only 100K data points
  - The right time to deliver analysis results is when the customer has enough data to make a decision – not before
Social Media Analytics requires **more** than Analytics

**Consumability**

- **Meaningful Social Media Metrics**
- **Driven by the Line of Business, not IT**
- **Deliver results to more than one employee**

**Capability**

- **Influencer Identification**
  - Network / Graph Analysis
- **Statistics**
  - Affinity patterns
- **Data Integration**
  - Author profile matching
  - Correlation with internal KPIs

**Text Analysis / NLP**

- Brands, Products, Product Features
- Sentiment, Mood, Emotion
- Author Location, Demographics, Behavior, Personality Traits
- Topic Clusters
Agenda

Social Media Analytics: Scope and Myths

What to measure in Social?

It depends on who you ask...

Analysis approaches and Challenges

Our Text Analysis Environment
Marketing: I want to enhance my reach in social media

I „lag“ behind my peers

What's tripadvisor doing here – and why am I missing?
Marketing: I want to find new sales channels for my retail brand

Proximity to a „Healthy Food“ store is seen as a plus
Start co-marketing activities with certain hotels to pull „healthy food“ shoppers to my store locations
Public Relations: I want to protect my brand reputation

Action

Check own supply chain to pro-actively avoid this problem
Prepare statement to clarify that *your* brand is not affected
Sales: I want to avoid customer churn or identify sales leads

Company One

Company Two

Identify engagement signals...

...before your competitor does

Thilo

CSR

Thilo

#Jump on over to

2 Alternatives

Thilo

Thilo

Social is only one view of the customer

Next step: Engage
Product Management: I want feedback on what consumers like/dislike around the competition

- Smell: 24.20%
- Stains: 7.40%
- Hygiene: 1.9%
- Palm oil: 1.5%
- Allergies: 1.2%
- Cold wash: 0.5%
- Discussions by Topic/Concern

"Smell" Related Discussions by Brand

- "Seriously, I definitely can’t recommend [brand A] to anyone with a sensitive nose! After 2 wash loads, the whole house smells – urgh!!"
- "Have again bought [brand E]. I just can’t take the smell of [brand A] any more and fortunately the bottle is almost empty. Somehow, my clothes start smelling kind of sour after a couple of days in the closet, when I’ve used [brand A] “
A Social Media Framework – defining and grouping Social Media KPIs

**Social Media Impact**
- Are we making the right investments in products, services, markets, campaigns, partners?
  - Measures:
    - Share of Voice
    - Reach
    - Sentiment

**Social Media Segmentation**
- Who are influencers and advocates?
- How can I identify sales leads from social media activity?
- Who are the customers I should reach out to?
  - Measures:
    - Geographics, Demographics
    - Users, Prospective Users
    - Advocates, Detractors
    - Influence

**Social Media Discovery**
- What new topics are emerging in social media?
  - Measures:
    - Topics
    - Participants
    - Sites

**Social Media Relationships**
- What is driving social media behaviour?
  - Measures:
    - Affinity
    - Association
    - Cause
How does it work? – “Inside” IBM Social Media Analytics 1.2

Welcome Page

Analytics Platform
- Topic Detection
- Text Analytics
- IBM BigInsights 1.3

Search Index

Influencer Identification

DB2 ESE 9.7

Cognos 10.1.1

E2E Orchestration

IBM Social Media Analytics

SSO

Multi-tenancy

Security

Blogs
Message boards
Twitter
News
Facebook
Reviews
Video

Project

Project

Project

Project

Project

Project
Agenda

Social Media Analytics: Scope and Myths
What to measure in Social?

Analysis approaches and Challenges

Our Text Analysis Environment
Text Analytics for Social Media Analysis

- **Goal**: extract information from what users write to
  - **Aggregate** information into meaningful statistics for end users, as well as extract
  - **Supporting evidence** for the statistics presented to users.

- **Types of information** include
  - **Sentiment**: are users writing positively or negatively about the product
  - **Demographics**: gender, age, family status, geographic information...
  - **Author “behavior”**: are they recommending or cautioning against the product, are they owners or potential buyers of the product....
Rules „vs.“ Machine Learning – the advantages of Rules

- High expressiveness: phenomena like comparisons are straightforward to express in rules
- Smaller amount of “human-coded” training data: Smaller adaption effort to new domains and languages
- Clear lineage:
  - If it doesn't work, we can understand why and can fix it quickly – even after several iterations
  - Transparency for our users
- More fine-grained: detect sentiment for a particular product, not a whole tweet
- Statistical approaches help us to build rules
The challenge in Social Media....

**OBI** in the German DIY world

**OBI** in the rest of the world

The **F50**

*Also the F50*
Capturing concepts (such as brands or products)

- Simple keywords are not enough – you sometimes need regular expressions to capture all variations

- Define concepts through include, context and exclude terms
  - Include terms “make up” the concept (including synonyms)
  - Context terms describe relevant contexts
  - Exclude terms rule out irrelevant meanings

- Examples:
  - Only match Obi when neither Wan, nor Kenobi, nor star wars are present (exclude)
  - Only match F50 when sports or running or adidas are also present (context)
Goal & Key challenge: **Only pick up sentiment that is relevant to a concept**  
*Yesterday I had a sports massage which was *wonderful*. So I went running with my new *Running XYs* – but got *blisters*

**Aggregate** the sentiment for each concept mention

- **Positive**: concept mention contains more positive than negative sentiment for the concept  
  *I've had a *Phone A* for a bit less than a month now and it's pretty *sweet*

- **Negative**: vice versa  
  *While I like my *Phone A* (despite it's many *flaws*) I am *not feeling all that confident* that I'll see Gingerbread on my device.*

- **Ambivalent**: equal amount of positive and negative sentiment  
  *The battery on *Phone A* is *good*, but the charging time *could be better*

- **Neutral**: concept mention doesn’t contain any sentiment around the concept  
  *On-device debug with *Phone A* *USB driver*. You need to install the device-specific *driver in addition to the SDK*
Configuring Sentiments for SMA Administrators

- **Add** or **remove** positive or negative sentiment terms & sentiment blockers

- **De-activate** sentiment terms
  - Term is kept in the sentiment list, but is not applied in snippets
  - Useful to keep terms „around“

- **No configuration** of grammar rules

![Sentiment configurations in IBM Cognos Consumer Insight](image-url)
Steps to detect sentiment – a rule-based approach

1. Detect positive and negative terms
   love, sweet – blisters, flaws

2. Remove terms that are covered by sentiment blockers
   issue vs. „January issue“

3. Apply syntax rules to determine negation, desires, questions...
   I’m confident vs. I’m not confident
   a problem vs. they solved the problem
   they improved their service vs. they should improve...
   They are good vs. Are they good?

4. Pick the sentiment phrases that are close to a concept
   - Can be based on source (e.g., blogs vs. reviews), proximity, grammatical constraints...
Example: **Concept-level** sentiment around the Sony Xperia Z

**iPhone 5s Outlasts iPhone 5 in Battery Tests**
10/6/13 10:00 PM

... This is almost two hours more than the iPhone 5, which lasted 8 hours and 42 minutes. However, Apple’s new iPhone failed to beat the talk time score of Sony’s latest camera smartphone, the Xperia Z1, which delivered the longest talk time, close to 27 hours. Web browsing test results for iPhone 5s ...

Show document  Language: English  Author: Sarmista Acharya  (Sarmista Acharya)  
Source: news  (International Business Times, India)

**RE: Welches Handy ist besser Samsung Galaxy s3 oder Sony Xperia z?????**
10/7/13 5:00 AM

... Hallo erstmal Ich habe jetzt das sony xperia z und davor hatte ich dass Samsung galaxy s3. Ich finde dass das sony xperia z viel besser ist ... Es hat eine bessere Kamera - Besser Grafik - Ist Wasser und staubdicht - schneller im Internet - Hat aber leider kein erweiterbaren Speicher ... Ich würde sagen das Sony Xperia z ist besser als dass Samsung galaxy s3. Das Samsung galaxy s3 besteht aus plastig und ist Nicht Grad das beste ...

No further content  Language: German  Author: babohi  (babohi)  
Source: boards  (Die beliebtesten Themen der Ratgeber-Community von Abnehmen bis Zähne)

**RE: Due an upgrade this month. Do I move away from the iPhone?**
10/7/13 9:00 PM

... Hi. I moved from an iPhone 4s to a Xperia Z1. I’m loving Android. Such a breath of fresh air ...

Show document  Language: English  Author: Vita  (Vita)  
Source: boards  (Overclockers UK Forums)

**RE: ****The Official Note III Thread****
10/6/13 7:00 PM

... ok so a note 3 it is ... I’ve had a good look at it in CFW today, Xperia Z1 looks poor in comparison, also the Z Ultra ... that’s one huge ridiculous sized device, almost as big as an iPad mini ...

Show document  Language: English  Author: maddness  (maddness)  
Source: boards  (Overclockers UK Forums)
Identifying author demographics

- **Gender**
  - Identify gender through cues from the author’s first name, the author’s nickname and the author content

- **Is author married (en, de, es, fr)**
  - Identified in author content through trigger terms and text analysis rules

  *Snippet:* Yes, Google owns a huge chunk of Motorola. This is precisely why my wife’s Motorola Droid RAZR MAXX is getting the new Android Jelly Bean update before my much more popular and better selling Samsung Galaxy S3.

- **Is author a parent (en, de, es, fr)**
  - Identified in author content through trigger terms and text analysis rules
  - Nicknames can be a good source of information as well („SuperMom2012“)

  *Snippet:* Just waiting for OTA JB and just rock that. I recall you’re on Speakout—my son is also with an unlocked Bell S3. I wonder if his S3 will get the OTA update through the Rogers network/Speakout?
Identifying author behavior (en, de, es, fr)

- Users of a certain product or service
  - What product features are relevant for them?

- Recommenders
  - E.g., authors mentioning „you should use X“

- Detractors
  - E.g. authors mentioning „stay away from X“

- Prospective users
  - Potential sales leads for 1:1 engagement
  - Identify sites where prospective users congregate
One road ahead: Deeper author-based insights
IBM researcher can decipher your personality from looking at 200 of your tweets

http://venturebeat.com/2013/10/08/ibm-researcher-can-decipher-your-personality-in-200-tweets/
Topic Detection in Social Media

- **Goal**: find „lists of keywords“ (=topics) that allow to „reconstruct“ a social media post through a combination of topics

- **Approach**: Non-Negative Matrix Factorization

- **Advantage over document clustering**: focus is on getting representative topic keywords, which helps the user to understand what he documents „are about“, not „perfect“ document clusters

- **Factorization problem**:

  \[
  \begin{align*}
  \text{Document-Term-Matrix } X & \quad \rightarrow \quad \text{Document-Topic-Matrix } W
  \\
  \text{term} & \rightarrow \text{K topics}
  \\
  \begin{pmatrix}
  0.1 & 0 & 0.15 & 0.2 & \ldots \\
  0.2 & 0.3 & 0.01 & 0 & \ldots \\
  0 & 0.1 & 0.02 & 0 & \ldots \\
  0 & 0.02 & 0.05 & 0.08 & \ldots \\
  0.01 & 0.07 & 0.08 & 0.09 & \ldots \\
  0.09 & 0.08 & 0.04 & 0.02 & \ldots \\
  0.08 & 0.02 & 0.01 & 0.02 & \ldots \\
  \ldots & \ldots & \ldots & \ldots & \ldots
  \end{pmatrix}
  \end{align*}
  \]

  \[
  \begin{align*}
  \text{Topic-Term-Matrix } H & \quad \rightarrow \quad \text{K topics}
  \\
  \text{term} & \rightarrow \text{K topics}
  \\
  \begin{pmatrix}
  0.1 & 0.2 & 0.15 & \ldots \\
  0.2 & 0.31 & 0.5 & \ldots \\
  0.6 & 0.1 & 0.1 & \ldots \\
  0.3 & 0.02 & 0.2 & \ldots \\
  0.01 & 0.07 & 0.09 & \ldots \\
  0.19 & 0.08 & 0.3 & \ldots \\
  0.4 & 0.02 & 0.02 & \ldots \\
  \ldots & \ldots & \ldots & \ldots
  \end{pmatrix}
  \end{align*}
  \]
Agenda

Social Media Analytics: Scope and Myths
What to measure in Social?
Analysis approaches and Challenges

Our Text Analysis Environment
Text Analytics Architecture – an “IBM view”

- Named entity extractors, Sentiment, Log analysis, Table extraction, etc.
- Scalable Runtime
- Tooling + UI
- Various levels of users
- Customizable Libraries
- Simplicity vs. expressiveness
- Extraction Language
- Tokenization, Lemmatization, POS tagging, Sentence boundary detection, Deep parsing, …
- Compiler + Optimizer
- Extraction Language
- Extraction Language
- Extraction Language
- Extraction Language
- Core Operations
- Execution Engine
- Core Operations
Architecture Implementation: IBM „System T“

- Execution Engine
- Compiler + Optimizer
- Extraction Language
- Generic Libraries
- Tooling + UI

- Eclipse tooling: editor, workflow analysis, pattern discovery, regex learner, profiler, ...
- NE libraries, Sentiment, informal-text normalizer, ...
- Cost based optimizer
- Scalable Embeddable Runtime
- 20+ languages currently supported
- AQL
AQL: A Declarative Language to Specify Extraction Patterns

```
create view Number as
extract regex /\d+(\.\d+)?/ on D.text as match
from Document D;
```

... Asia-Pacific revenues increased 7 percent (5 percent, adjusting for currency) to $4.8 billion. OEM revenues were $1.0 billion, down 3 percent compared with the 2005 fourth quarter.

Choice of SQL-like syntax for AQL motivated by wider adoption of SQL
create view Unit as
extract dictionary 'UnitDict' 
    with flags 'IgnoreCase'
on D.text as match
from Document D;

...Asia-Pacific revenues increased 7 percent (5 percent, adjusting for currency) to $4.8 billion. OEM revenues were $1.0 billion, down 3 percent compared with the 2005 fourth quarter.

...
AQL example: matching sequences

```
create view AmountWithUnit as
extract pattern <N.match> <U.match>
as match
from Number N, Unit U;
```

... 
Asia-Pacific revenues increased 7 percent (5 percent, adjusting for currency) to $4.8 billion. OEM revenues were $1.0 billion, down 3 percent compared with the 2005 fourth quarter. 
...
AQL expressiveness

- Similar to the standard relational model used by SQL databases like DB2
- All data in AQL is stored in **tuples**: data records of one or more columns/fields
- Basic extraction constructs
  - EXTRACT statement
  - Regular expression
  - Dictionaries
  - Sequence pattern
- Relational-style constructs
  - SELECT
  - JOIN
  - UNION ALL and MINUS statements
- Aggregation operators
  - CONSOLIDATE
  - BLOCK
AQL Eclipse Tools Overview

Ease of Programming

AQL Editor: syntax highlighting, auto-complete, hyperlink navigation
Result Viewer: visualize/compare/evaluate
Explain: show how each result was generated
Workflow UI: enable novice users to become experts in a short time

Automatic Discovery

Pattern Discovery: identify patterns in the data
Regex Generator: generate regular expressions from examples

Performance Tuning

Profiler: identify performance bottlenecks to be hand tuned
Many different possible ways to execute a given set of AQL statements

The SystemT Optimizer chooses a good plan from among the alternatives

Employs multiple techniques
  - AQL rewrite rules
  - Cost-based optimization
  - Global plan rewrite rules
What is an operator graph?

- An extractor plan is a graph of operators.
- Operator: A module that performs a specific task, like identifying matches of a regular expression on a string.
- The output of one operator becomes the input of another.
create view Number as
extract regex /\d+/ on between 1 and 1 tokens in D.text as match from Document D;

create view Unit as
extract dictionary UnitDict on D.text as match from Document D;

create view AmountWithUnit as
extract pattern <N.match> <U.match> return group 0 as match from Number N, Unit U;
Example Optimization: Conditional Evaluation (CE)

- Leverage document-at-a-time processing
- Don’t evaluate the inner operand of a join if the outer has no results

Don’t evaluate this Regex when there are no dictionary matches.
Profiler Output: “Hot” Views

---

Top 25 Views by Execution Time:
---

<table>
<thead>
<tr>
<th>View Name</th>
<th>Samples</th>
<th>Seconds</th>
<th>% of Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>DateISO</td>
<td>665</td>
<td>0.79</td>
<td>1.30</td>
</tr>
<tr>
<td>DateISOExtended</td>
<td>677</td>
<td>0.81</td>
<td>1.32</td>
</tr>
<tr>
<td>CodeCharNumSymBaseUnfiltered</td>
<td>2395</td>
<td>2.86</td>
<td>4.67</td>
</tr>
<tr>
<td>DateNormalized</td>
<td>2397</td>
<td>2.86</td>
<td>4.68</td>
</tr>
<tr>
<td>Time4Follows3\u2761subquery1</td>
<td>2871</td>
<td>3.43</td>
<td>5.60</td>
</tr>
</tbody>
</table>

- Views whose compiled plans are responsible for the largest fraction of execution time
- The view at the bottom of the list is the most expensive
Summary

- Social Media Analytics covers Content, People and Relationships found in Social Media Data

- Social Media Analytics is relevant to several enterprise users – across PR, Marketing, Sales, Product management, Brand Strategy
  - Different KPIs are relevant to different people – it’s not always about sentiment
  - Requires more than just a set of „technology blocks“

- One key technology for Social Media Analysis is Text Analysis
  - Content-level insights like brands, products, features, sentiment
  - Author-level insights like location, demographics, behavior

- We’re ALWAYS interested in interns:
Additional Information

- IBM Social Media Analytics

- IBM Social Media Analytics product videos
  http://ibmtdemo.edgesuite.net/software/analytics/cognos/videos/HTVs/sma-1-2/index.html

- Integrating social data with BI and Predictive Analytics
  http://www-01.ibm.com/support/docview.wss?uid=swg27038638