Text analytics and new visualization techniques

Richard Brath
rbrath@unchartedsoftware.com
brathr@lsbu.ac.uk
@rkbrath
richardbrath.wordpress.com

Scott Langevin
slangevin@unchartedsoftware.com
blog.unchartedsoftware.com
@slangevi
Most Common Text Visualization

Tag cloud via wordle.com. Note color is random, location is random and size is ambiguous – does Alice occur 10x or 100x more than Queen?
1. THE 500 YEAR CONSPIRACY
against text visualization
Genealogical tree from late 1400’s. Note graph nodes use of image (people, shield) or text, where text may be black, red or start with a red initial. The nodes can vary in size, color, or shape (circle, crescent, shield). Textual commentary is intertwined throughout.

1439 Johannes Gutenberg changes everything

1573: Image separates from text. From William Bullein's A Dialogue... Against the Fever Pestilence. Author photo from Bodleian exhibition "Shakespeare's Dead".
Diderot’s Encyclopedia has great illustrations of various occupations – all neatly labeled, but the viewer has to cross-reference the text to understand.

The Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers, par une Société de Gens de lettres was published under the direction of Diderot and d’Alembert, with 17 volumes of text and 11 volumes of plates between 1751 and 1772. Containing 74,000 articles written by more than 130 contributors, the Encyclopédie was a massive reference work for the arts and sciences, a "machine de guerre" which served to propagate the ideas of the French Enlightenment. The impact of the Encyclopédie was enormous. Through its attempt to classify learning and to open all domains of human activity to its readers, the Encyclopédie was one of the most important intellectual and social developments of its time.

**INDIGESTION ET MANIOC.**


**ECONOMIE RUSTIQUE.**

La vignes est illustrée par une invention, A. le vignes, B. le vignes, C. le vignes, D. le vignes, E. le vignes, F. le vignes, G. le vignes, H. le vignes, I. le vignes, J. le vignes, K. le vignes, L. le vignes, M. le vignes, N. le vignes, O. le vignes, P. le vignes, Q. le vignes, R. le vignes, S. le vignes, T. le vignes, U. le vignes, V. le vignes, W. le vignes, X. le vignes, Y. le vignes, Z. le vignes.
1920
Same for charts

1919 book explaining charts. Labels and notes are pushed to the periphery of the chart. (Willard C. Brinton, Graphic Methods for Presenting Facts, 1919)
Modern visualizations in news. Note how there is a lot of text – but moved out and round the charts.

---


Pyroman


https://digitalcomicmuseum.com?did=13773&page=27

Outside the Laboratory...

Der Boss said the laboratory is developing an important war secret. Heinrich is up to us to find out more about it!

Jai! We will...wait! Here come two of der laboratory workers!

Sure...the new radio activator turns any gas into a solid...instantaneously!

What the...!

That evening...

Up with your hands—quick! Hang, grab der radio-activator!

What the—!

In the Central Tech Laboratory...

It's a great achievement, Dick! We can now actually swing it into battle against the forces of evil!

It will speed up war construction tremendously, by eliminating such foolishness as—

Here's where we make a try at it, anyway!

Here is where I should kill you...but we need you to show us how to work der machine!—

At that moment...

Verdammt! We will have to get away fast! Come!

Dick! What?—

Boss! Der police! Der must have heard der fighting!
2. Bringing **TEXT** back into **VISUALIZATION**
Plain Dots

- Simple scatterplot
- Macro pattern e.g. crescent shape
- Also color and size set to population

Image: Birth rate vs. Death rate per country, with size and color indicating population; and diagonal lines indicating birth to death ratio. Note visible crescent shape. Image by author.
Alphanumeric Points

- Same scatterplot
- 2-letter ISO codes added
- More information, more insight

Image: 2 letter country ISO codes act as mnemonics. Note codes upper left: LV, UA, RU, EE – Latvia, Ukraine, Russia, Estonia, etc., all below population replacement rate. Image by author.
Better than label and bar: See artist, title and first line.

Simple label has problems:
Artist as label doesn’t show the song. Song as label is ambiguous (e.g. who’s version of Silent Night?). Label using artist + song is too long and leaves little space for bar (e.g. see entry for Elton John).

First line provides rich context: maybe you don’t know the Wabash Cannonball, but with the first line you know what it’s about. Images by author.

---

**Best-Selling Singles**

* more than 10 million copies, sorted by release date

<table>
<thead>
<tr>
<th>Artist</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bing Crosby</td>
<td>&quot;Silent Night&quot;</td>
</tr>
<tr>
<td>The Ink Spots</td>
<td>&quot;Night and Day&quot;</td>
</tr>
<tr>
<td>Roy Acuff</td>
<td>&quot;Babys Ball&quot;</td>
</tr>
<tr>
<td>The Mills Brothers</td>
<td>&quot;Paper Doll&quot;</td>
</tr>
<tr>
<td>Gene Autry</td>
<td>&quot;Rudolph the Red-Nosed Reindeer&quot;</td>
</tr>
<tr>
<td>Patti Page</td>
<td>&quot;Tennessee Waltz&quot;</td>
</tr>
<tr>
<td>Bill Haley</td>
<td>&quot;Rock Around the Clock&quot;</td>
</tr>
<tr>
<td>The Penguins</td>
<td>&quot;Earth Angel&quot;</td>
</tr>
<tr>
<td>Elvis Presley</td>
<td>&quot;Hound Dog&quot;</td>
</tr>
<tr>
<td>Domenico Modugno</td>
<td>&quot;Nel blu dipinto di blu (Volare)&quot;</td>
</tr>
<tr>
<td>Elvis Presley</td>
<td>&quot;It's Now or Never&quot;</td>
</tr>
<tr>
<td>Kyu Sakamoto</td>
<td>&quot;Sukiyaki&quot;</td>
</tr>
<tr>
<td>The Beatles</td>
<td>&quot;I Want to Hold Your Hand&quot;</td>
</tr>
<tr>
<td>The Monkees</td>
<td>&quot;I'm a Believer&quot;</td>
</tr>
<tr>
<td>Proc Harol</td>
<td>&quot;A White Shade of Pale&quot;</td>
</tr>
<tr>
<td>Monge Jorry</td>
<td>&quot;In the Summertime&quot;</td>
</tr>
<tr>
<td>Carl Douglas</td>
<td>&quot;Kung Fu Fighting&quot;</td>
</tr>
<tr>
<td>George McCrae</td>
<td>&quot;Rock Your Baby&quot;</td>
</tr>
<tr>
<td>Roger Whittaker</td>
<td>&quot;The Last Farewell&quot;</td>
</tr>
<tr>
<td>ABBA</td>
<td>&quot;Fernando&quot;</td>
</tr>
<tr>
<td>John Travolta</td>
<td>&quot;Olivia Newton-John&quot;</td>
</tr>
<tr>
<td>Gloria Gaynor</td>
<td>&quot;I Will Survive&quot;</td>
</tr>
<tr>
<td>The Trio</td>
<td>&quot;Da Da Da&quot;</td>
</tr>
<tr>
<td>Bryan Adams</td>
<td>&quot;(Everything I Do) I Do It For You&quot;</td>
</tr>
<tr>
<td>Scorpions</td>
<td>&quot;Wind of Change&quot;</td>
</tr>
<tr>
<td>Whitney Houston</td>
<td>&quot;I Will Always Love You&quot;</td>
</tr>
<tr>
<td>Andrea Bocelli &amp; Sarah Brightman</td>
<td>&quot;Time to Say Goodbye&quot;</td>
</tr>
<tr>
<td>Toni Braxton</td>
<td>&quot;Un-Break My Heart&quot;</td>
</tr>
<tr>
<td>Elton John</td>
<td>&quot;Candle in the Wind&quot;</td>
</tr>
<tr>
<td>Celine Dion</td>
<td>&quot;My Heart Will Go On&quot;</td>
</tr>
<tr>
<td>Cher</td>
<td>&quot;Believe!&quot;</td>
</tr>
<tr>
<td>Panjabi MC</td>
<td>&quot;Mundian Toh Bach Ke&quot;</td>
</tr>
<tr>
<td>Britney Spears</td>
<td>&quot;Oops...I Did It Again&quot;</td>
</tr>
<tr>
<td>Celine Dion &amp; Andrea Bocelli</td>
<td>&quot;Prayer&quot;</td>
</tr>
</tbody>
</table>

---

Best-Selling Singles more than 10 million copies, sorted by release date

Bing Crosby -- "Silent Night" - Silent night, Holy night, . . .
The Ink Spots -- "If I Care, I Won't Care" - If I didn't care, would I feel more? This isn't the time to say...
Bing Crosby -- "White Christmas" - I'm dreaming of a white Christmas, Just like the ...u.
Roy Acuff -- "Nabash Cannon Mills Brothers -- "Paper Doll" - I'm gonna buy a paper doll that can call my name...
Gene Autry -- "Rudolph the Red-Nosed Reindeer" -- You'll never see a sleigh ride finer...
Patti Page -- "Tennessee Waltz" - When I was dancing with you,...
Bill Haley & His Comets -- "Rock Around the Clock" - Rock, rock, roll on...
The Penguins -- "Earth Angel" - Little white 地方
Elvis Presley -- "Hound Dog" - Hey baby I'm here to knock you over...
Frankie Avalon & Annette -- "Donna"
Domenico Modugno -- " Nel blu dipinto di blu (Volare)"
Elvis Presley -- "It's Now or Never" - It's now or never...
Kyu Sakamoto -- "Sukiyaki" - Let's look up as we walk...
The Beatles -- "I Want To Hold Your Hand" - Oh yeah...
The Monkees -- "I'm a Believer" - Love was a made of Pale...
Nuno Gomes -- "In the Summertime" - In the summertime when the weather is hot...
Middle of the Road -- "Chirpy Cheep Cheep"
Carl Douglas -- "Kung Fu Fighting" - Now here it is...
George McCrae -- "Rock Your Baby" - Sexy mama, Woman...
Roger Whittaker -- "The Last ABBA -- "Fernando" - Can you...
John Travolta & Olivia Newton-John -- "You're the One That I Want" - There's a feeling that you're not...
Gloria Gaynor -- "I Will Survive" - At first I was scared...
Yakolev Trio -- "Da Da Da" - Aha aha aha, Aha, aha aha, Was Band Aid -- "Do They Know It's Christmas?" - It's Christmas for Africa...
Bryan Adams -- "Everything I Do" - It's for you...
Scorpions -- "Wind of Change" - I feel the Moskva...
Whitney Houston -- "I Will Always Love You" - If I should leave this world today...
Macarena -- "Macarena" - And she says...
Andrea Bocelli & Sarah Brightman -- "Time to Say Goodbye" - My heart...
Toni Braxton -- "Un-Break My Heart" - Don't leave me...
Elton John -- "Candle in the Wind 97" - Something About the Way You Look Tonight - Go goodbye...
Celine Dion -- "My Heart Will Go On" - Every night in my dreams...
Celine Dion -- "My Heart Will Go On" - Every night in my dreams...
"I Will Always Love You" - If I should leave this world today...
Panchita Mc -- "Mundian To De"
Britney Spears -- "Baby Baby Baby Baby Baby Baby" - I'm a woman...
Celine Dion & Andrea Bocelli -- "The Prayer" - I pray...

Bar Chart of Text

Plus content analysis:

Is “love” the top word in the top 40?

Bar Chart of Text

No: “Christmas” beats love (but occurs in only 2 songs)

Best-Selling Singles more than 10 million copies, sorted by release date

<table>
<thead>
<tr>
<th>Song Title</th>
<th>Artist(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silent Night</td>
<td>Bing Crosby</td>
</tr>
<tr>
<td>I Care</td>
<td>Bing Crosby</td>
</tr>
<tr>
<td>Be True</td>
<td>Bing Crosby</td>
</tr>
<tr>
<td>White Christmas</td>
<td>Bing Crosby</td>
</tr>
<tr>
<td>Roy Acuff</td>
<td>nabash Canno</td>
</tr>
<tr>
<td>Paper Doll</td>
<td>Mills Brothers</td>
</tr>
<tr>
<td>Rudolph the Red-Nosed Reindeer</td>
<td>Gene Autry</td>
</tr>
<tr>
<td>Tennessee Waltz</td>
<td>Patti Page</td>
</tr>
<tr>
<td>Rock Around the Clock</td>
<td>Bill Haley &amp; His Comets</td>
</tr>
<tr>
<td>Earth Angel</td>
<td>The Penguins</td>
</tr>
<tr>
<td>Hound Dog</td>
<td>Elvis Presley</td>
</tr>
<tr>
<td>Dia:Beach</td>
<td>Domenico Modugno</td>
</tr>
<tr>
<td>Nel blu dipinto di blu</td>
<td>Elvis Presley</td>
</tr>
<tr>
<td>It's Now or Never</td>
<td>Kyu Sakamoto</td>
</tr>
<tr>
<td>Sukiyaki</td>
<td>The Beatles</td>
</tr>
<tr>
<td>I Want to Hold Your Hand</td>
<td>The Monkees</td>
</tr>
<tr>
<td>I'm a Believer</td>
<td>Procol Harum</td>
</tr>
<tr>
<td>In the Summertime</td>
<td>Mungo Jerry</td>
</tr>
<tr>
<td>I Don't Care</td>
<td>Middle of the Road</td>
</tr>
<tr>
<td>Kung Fu Fighting</td>
<td>Carl Douglas</td>
</tr>
<tr>
<td>Rock Your Baby</td>
<td>George McCrae</td>
</tr>
<tr>
<td>Farewell</td>
<td>Roger Whitaker</td>
</tr>
<tr>
<td>Can You Feel it</td>
<td>The Last Abba</td>
</tr>
<tr>
<td>Someone is Waiting</td>
<td>John Travolta &amp; Olivia Newton-John</td>
</tr>
<tr>
<td>I Will Survive</td>
<td>Gloria Gaynor</td>
</tr>
<tr>
<td>Da Da Da</td>
<td>Trubio</td>
</tr>
<tr>
<td>Do They Know It's Christmas</td>
<td>Band Aid</td>
</tr>
<tr>
<td>America</td>
<td>USA for Africa</td>
</tr>
<tr>
<td>lambada</td>
<td>Chorando se foi</td>
</tr>
<tr>
<td>Everything I Do Do You</td>
<td>Bryan Adams</td>
</tr>
<tr>
<td>Wind of Change</td>
<td>Scorpions</td>
</tr>
<tr>
<td>I Will Always Love You</td>
<td>Whitney Houston</td>
</tr>
<tr>
<td>White Macadam</td>
<td>Andrea Bocelli &amp; Sarah Brightman</td>
</tr>
<tr>
<td>Time to Say Goodbye</td>
<td>Toni Braxton</td>
</tr>
<tr>
<td>Un-Break My Heart</td>
<td>Elton John</td>
</tr>
<tr>
<td>Candle in the Wind 1997!</td>
<td>Celine Dion</td>
</tr>
<tr>
<td>My Heart Will Go On</td>
<td>cher - Believe</td>
</tr>
<tr>
<td>Nothing</td>
<td>Panjabi MC</td>
</tr>
<tr>
<td>Mandarin To</td>
<td>Britney Spears</td>
</tr>
<tr>
<td>Baby</td>
<td>Celine Dion &amp; Andrea Brightman</td>
</tr>
</tbody>
</table>


Best Selling Singles more than 10 million copies, sorted by release date

0 10 20 30 40 50m

Round you virgin mother, child, and Holy infant
Save us to this hour, from all our harms affright.
All is calm, all is bright!

The endless germs of flowing mountain to the
ows cannot steal, And then
Cupid and Donder and Blitzen
I happened to see, I intro
seven o’clock, eight-o’clock
time, I'm just a fool,
dog, scratchin’ all the
they say, cause ford
tears, And in the hustle
tomorrow, Tonight will be
ting day, that lonely
thing that I say that something, I want
for someone else but not
to love, Love was out to
cartwheels across the floor.
you can stretch right up and touch the sky, cause ford

Bar Chart of Text

No: “Christmas”

beats love
(but occurs in only 2 songs)
Bar Chart of Text

And “baby” beats both in counts and songs

Best-Selling Singles more than 10 million copies, sorted by release date

<table>
<thead>
<tr>
<th>Artist</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bing Crosby</td>
<td>&quot;Silent Night&quot;</td>
</tr>
<tr>
<td>The Everly Brothers</td>
<td>&quot;Let It Snow, Let It Be&quot;</td>
</tr>
<tr>
<td>Bing Crosby</td>
<td>&quot;White Christmas&quot;</td>
</tr>
<tr>
<td>Roy Acuff</td>
<td>&quot;Nashville Cannonball&quot;</td>
</tr>
<tr>
<td>Willie Nelson</td>
<td>&quot;Rock Around the Clock&quot;</td>
</tr>
<tr>
<td>The Penguins</td>
<td>&quot;Earth Angel&quot;</td>
</tr>
<tr>
<td>Elvis Presley</td>
<td>&quot;Hound Dog&quot;</td>
</tr>
<tr>
<td>Domenico Modugno</td>
<td>&quot;Nel Blu Dipinto Di Blu&quot;</td>
</tr>
<tr>
<td>Elvis Elvis</td>
<td>&quot;It's Now or Never&quot;</td>
</tr>
<tr>
<td>Kyu Sakamoto</td>
<td>&quot;Sukiyaki&quot;</td>
</tr>
<tr>
<td>The Beatles</td>
<td>&quot;I Want to Hold Your Hand&quot;</td>
</tr>
<tr>
<td>The Monkees</td>
<td>&quot;I'm a Believer&quot;</td>
</tr>
<tr>
<td>Procol Harum</td>
<td>&quot;A Whiter Shade of Pale&quot;</td>
</tr>
<tr>
<td>The Middle Road of the Chills</td>
<td>&quot;Carl Douglas&quot;</td>
</tr>
<tr>
<td>Mark Mulcahy</td>
<td>&quot;Kung Fu Fighting&quot;</td>
</tr>
<tr>
<td>George McCrae</td>
<td>&quot;Baby, Baby, Baby&quot;</td>
</tr>
<tr>
<td>Roger Whittaker</td>
<td>&quot;The Last Farewell&quot;</td>
</tr>
<tr>
<td>ABBA</td>
<td>&quot;Fernando&quot;</td>
</tr>
<tr>
<td>John Travolta &amp; Olivia Newton-John</td>
<td>&quot;The One&quot;</td>
</tr>
<tr>
<td>Gloria Gaynor</td>
<td>&quot;I Will Survive&quot;</td>
</tr>
<tr>
<td>The Jackson 5</td>
<td>&quot;Da Da Da&quot;</td>
</tr>
<tr>
<td>Bryan Adams</td>
<td>&quot;Do They Know It's Christmas?&quot;</td>
</tr>
<tr>
<td>Whitney Houston</td>
<td>&quot;I Will Always Love You&quot;</td>
</tr>
<tr>
<td>Andrew Bocelli &amp; Sarah Brightman</td>
<td>&quot;Time to Say Goodbye&quot;</td>
</tr>
<tr>
<td>Elton John</td>
<td>&quot;Candle in the Wind&quot;</td>
</tr>
<tr>
<td>Celine Dion</td>
<td>&quot;My Heart Will Go On&quot;</td>
</tr>
<tr>
<td>Panjabi MC</td>
<td>&quot;Mundian To Bach Ke&quot;</td>
</tr>
<tr>
<td>Britney Spears</td>
<td>&quot;Baby One More Time&quot;</td>
</tr>
<tr>
<td>Celine Dion &amp; Andrea Bocelli</td>
<td>&quot;The Prayer&quot;</td>
</tr>
</tbody>
</table>

3. But what about **BIG DATA TEXT ANALYTICS**?
Text Analytics 101

Typical “Natural Language Processing” Pipeline

Raw Text → Preprocessing → Numerical Representation → Analysis

https://en.wikipedia.org/wiki/Parse_tree
http://brandonrose.org/clustering
## Preprocesing Raw Text

<table>
<thead>
<tr>
<th>Step</th>
<th>Operation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokenize</td>
<td>words</td>
<td>“Hello, I’m Dr. Jones.”</td>
</tr>
<tr>
<td></td>
<td>sentences</td>
<td>['Hello', 'I', 'm', 'Dr', 'Jones']</td>
</tr>
<tr>
<td>Normalize</td>
<td>stem</td>
<td>ponies</td>
</tr>
<tr>
<td></td>
<td>lemmatize</td>
<td>ponies</td>
</tr>
<tr>
<td></td>
<td>lowercase</td>
<td>Ponies</td>
</tr>
<tr>
<td>Remove Noise</td>
<td>stop words</td>
<td>John’s car is red, right?</td>
</tr>
<tr>
<td></td>
<td>punctuation</td>
<td>John’s car is red, right?</td>
</tr>
</tbody>
</table>
Document Representation

Counting
- word count
- sentence count
- syllable count

TFIDF
\[
\text{TFIDF}_t^d = \frac{1}{\# \text{docs term } t \text{ appears in}} \times \frac{\# \text{times term } t \text{ is in doc } d}{\text{count}_t}
\]

<table>
<thead>
<tr>
<th></th>
<th>king</th>
<th>queen</th>
<th>man</th>
<th>woman</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>32</td>
<td>147</td>
<td>14</td>
<td>83</td>
</tr>
<tr>
<td>TFIDF</td>
<td>0.001</td>
<td>0.097</td>
<td>0.051</td>
<td>0.147</td>
</tr>
</tbody>
</table>

Word Embedding
Text Vectors

Document Representation

Part-of-Speech tagging

['And, 'now', 'for', 'something', 'completely', 'different']

'CC' 'RB' 'IN' 'NN' 'RB' 'JJ'

Chunking / Phrase extraction

Dependency parse tree


https://en.wikibooks.org/wiki/LaTeX/Linguistics#/

**Text Analysis**

**Topic Modeling**

Using statistical features to discover abstract ‘topics’ that occur in a collection of documents

‘topic’ is a probability distribution of words

- **Topic 1**: algorithm, computer, data, ...
- **Topic 2**: gene, data, bio, ...
- **Topic 3**: math, equation, matrix, ...
- **Topic 4**: chemical, carbon, water, ...

---


Brath and Jonker. Graph Analysis and Visualization. 2015. Wiley.  
## Other NLP Tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coreference resolution</td>
<td>&quot;Scott took a flight. He said it was very crowded&quot;</td>
</tr>
<tr>
<td>Text Classification</td>
<td>Spam / Not Spam</td>
</tr>
<tr>
<td>Text Similarity</td>
<td>(definitely, definately) = 1</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>Strata was great this year! :) = POSITIVE</td>
</tr>
<tr>
<td>Entity extraction</td>
<td>New York is a metropolis → ('New York', 'city')</td>
</tr>
<tr>
<td>Text Summarization</td>
<td>Mostly harmless</td>
</tr>
<tr>
<td>Machine Translation</td>
<td>Mostly harmless -&gt; ほとんど無害</td>
</tr>
<tr>
<td>Natural Language Generation</td>
<td>&quot;Sounds good, talk to you then&quot;</td>
</tr>
</tbody>
</table>
4. TEXT ANALYTICS & VIZ together!
Character Traits

Who are the characters?
What are their traits?

Tag cloud showing frequency of words in *Grimms' Fairy Tales*, a collection of 50+ short stories. Tag clouds reveals some characters, e.g. king, princess, fox; but doesn’t provide any insight into those characters. (via wordle.net)
NLP to find Traits

1. Extract the characters
2. Extract the adjectives near the characters
3. Visualize

e.g. This example, one sentence from Grimms’ Fairy Tales.

When you come to the old king, and he asks for the beautiful princess, you must say, “Here she is!”

When you come to the old **king**, and he asks for the beautiful **princess**, you must say, “Here she is!”

When you come to the **old king**, and he asks for the **beautiful princess**, you must say, “Here she is!”
### Characters from Grimms’ Fairy Tales with Associated Adjectives Weighted by Frequency

<table>
<thead>
<tr>
<th>Character</th>
<th>List of adjectives, weighted by frequency: 0 3 4-5 6-9 10+</th>
</tr>
</thead>
<tbody>
<tr>
<td>bird</td>
<td>beautiful, splendid, open, wooden, like, hanging</td>
</tr>
<tr>
<td>cat</td>
<td>little, one, long</td>
</tr>
<tr>
<td>fox</td>
<td>old, dead, young, first, fast</td>
</tr>
<tr>
<td>gretel</td>
<td>little, poor, good</td>
</tr>
<tr>
<td>hansel</td>
<td>little, like, fat</td>
</tr>
<tr>
<td>bride</td>
<td>false, true, first, right, real</td>
</tr>
<tr>
<td>king</td>
<td>old, great, one, young, three, angry, beautiful, married, like, third</td>
</tr>
<tr>
<td>princess</td>
<td>beautiful, young, last, dear, enchanted, strange, true, third, free</td>
</tr>
<tr>
<td>girl</td>
<td>little, poor, lazy, pretty, young, dead, beautiful, silly</td>
</tr>
<tr>
<td>queen</td>
<td>beautiful, late, little, far</td>
</tr>
<tr>
<td>wife</td>
<td>standing, married, next, poor, two, beautiful, new, one, dear, true</td>
</tr>
<tr>
<td>witch</td>
<td>old, wicked</td>
</tr>
</tbody>
</table>

Approximate number of unique adjectives: 10
**CEO’s Use of Emotion Words During Earnings Calls, weighted by frequency**

<table>
<thead>
<tr>
<th>CEO</th>
<th>List of adjectives, weighted by frequency:</th>
<th>2 3-4 5-7 8-11 12+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carl Bass</td>
<td>pleased, good</td>
<td></td>
</tr>
<tr>
<td>Carl Trowell</td>
<td>demand, lower, good, debt, improve, cash, older, grow, tender, retirement, late, termination</td>
<td></td>
</tr>
<tr>
<td>Carlos A. Rodriguez</td>
<td>grow, confidence, pleased, success, excited, proud, progress</td>
<td></td>
</tr>
<tr>
<td>Carol M. Meyrowitz</td>
<td>pleased, confident, share, margin, grow, powerful, organization, excellent, shopping, terrific, passionate, gain</td>
<td></td>
</tr>
<tr>
<td>Charles Alutto</td>
<td>good, kind, waste, cash, excited, lines, share, safe, hazardous, versus, fee, pleased, resources, tough, respect</td>
<td></td>
</tr>
<tr>
<td>Charles B. Stanley</td>
<td>completion, improvement, lower, gross, progress, reject, rejection</td>
<td></td>
</tr>
<tr>
<td>Charles Bunch</td>
<td>improvement, lower, share, cash, versus, margin, unfavorable, gross, good, aggressive, tax, income, kind</td>
<td></td>
</tr>
<tr>
<td>Charlie Scharf</td>
<td>kind, good, share, change, lower, grow, pay, pleased, cross, case, versus, helpful, terrific, progress, excited, challenge</td>
<td></td>
</tr>
<tr>
<td>Chris Connor</td>
<td>share, kind, margin, good, pretty, demand, gross, john, confidence, versus, income, cash, bad, interior, rent, useful</td>
<td></td>
</tr>
<tr>
<td>Chris Sinclair</td>
<td>good, progress, pretty, hot, content, grow, share, change, organization, improvement, gross, finally, kind</td>
<td></td>
</tr>
<tr>
<td>Christopher Crane</td>
<td>good, create, late, hearing, grow, successful, pleased, share, advocacy, regulatory, clean, majority, kind, debt</td>
<td></td>
</tr>
<tr>
<td>Chuck Bunch</td>
<td>good, improvement, share, unfavorable, versus, cash, lower, excellent, demand, respect, difficult, debt, income</td>
<td></td>
</tr>
<tr>
<td>Chuck Jones</td>
<td>case, cash, good, kind, improvement, progress, finally, cutting, inflation, court, improve, share, risk, electric</td>
<td></td>
</tr>
<tr>
<td>Clay Williams</td>
<td>lower, demand, pretty, kind, good, excluding, improve, completion, versus, share, grow, deflation, tough, difficult</td>
<td></td>
</tr>
<tr>
<td>Clifton Penoble</td>
<td>margin, lower, income, gross, finally, excited, share, pretty</td>
<td></td>
</tr>
<tr>
<td>Craig A. Menear</td>
<td>share, pleased</td>
<td></td>
</tr>
</tbody>
</table>

Approximate number of unique adjectives: 0, 5, 10
CEO’s Use of Emotion Words During Earnings Calls, weighted by frequency

List of adjectives, weighted by frequency: 2 3-4 5-7 8-11 12+

- Andrew P. Wilson: battlefield star console entertainment dragon fun ultimate
- Bruce Broussard: star tax enforce good progress successful
- David Hallal: disease complement devastating progress therapeutic success majority hospital
- Fabrizio Freda: beauty exciting prestige successful
- Gracia Martore: love content respect terrific apologize lagging
- J.P. Bilbrey: chocolate shanghai lower unfavorable force
- Jim Tisch: diamond cash income special pay doubt lose
- John D. Idol: luxury demand shopping watch bottom
- John Ferriola: pig improvement cash lower challenge
- John Lechleiter: cancer disease deal regulatory tax therapeutic
- Kelly King: mortgage risk income tough money fee
- Larry D. Young: punch case music football love passion
- Les Moonves: content deal bang success terrific confident scorpion late
- Mark Parker: basketball demand excited elite offense discipline
- Mark R. Hunter: beer tax lower unfavorable loss organization banquet blue mad
- Omid Kordestani: shopping holiday progress pay music watch friction
- Richard J. Harshman: demand fell defect excellent labor wages aggressive falling
- Steven Temares: sue shopping baby improve excited exciting share success pleased fulfillment
- Thomas Kennedy: missile defense demand government award bottom threat

Focusing on unique terms (TFIDF) characterizes the companies.

Electronic Arts

Estée Lauder

Alexion Pharmaceuticals

Loews

Approximate number of unique adjectives: 10
5. **STICKY** social media
Superbowl 50 Retweets

Uncharted Flare Chart. Vertical axis is number of retweets. Lines are individual tweets. Some tweets flare up quickly then flatten, some go up, stutter, then go up more, etc..
Superbowl 50 Retweets

Note: some retweet values are offscreen

Uncharted Flare Chart. Betty White goes up flattens a bit, then up more, repeats. Live version: http://unchartedsoftware.github.io/salt-core/demos/superbowl-retweets/
Uncharted Flare Chart. Coldplay has a popular tweet at 8:45 pm which then flattens out.
Uncharted Flare Chart. Esurance has a retweet campaign which pops up quickly but starts to flatten after 20 minutes.
Uncharted Flare Chart. This tweet from the International Space Station doesn’t flare up, but keeps getting more and more retweets. Note how only one tweet can be examined at a time.
Unemployment Rate as Percent Civilian Labour Force

Similar problem: many lines on a line chart are difficult to disambiguate. Only way to disambiguate is to interact with each line individually. Slow!

Unemployment Rate as Percent Civilian Labour Force
2000-2014 by Country. Each line is labeled with microtext in multiple languages.

Top Hashtags

Sept 2017

Apple event generates flares.
Top Hashtags

Sept 2017

Hurricanes are popular hashtags, even misspellings.
Top Hurricane Retweets

Sep. 6-14, 2017

Comedic flare...

@PopeQuanPaul – Oh we good. Trump won't let Jose in the US.
@jdrudd - Hurricane #Irma now a category 5 and it's a beast. It's larger than the state of Ohio, if that helps put it in perspective.
@PascoSheriff_ To clarify, DO NOT shoot weapons @ #Irma. You won't make it turn around; it will have very dangerous side effects.
So What

Sticky is about a lot of things
• number retweets
• longevity
• reach across communities
• who’s saying it
• what they’re saying

And…
• geographic reach
• topics
• demographic segmentation
• message impact
• drift over time
Salt

TILE-BASED VISUAL ANALYTICS
• Hierarchical data tiling using cluster computing.
• Interactive on-demand image tile generation.
• Layers of raw data and derivative analytics.

OPEN SOURCE
• Saltlib.com
• Built on Apache Spark, Hadoop and ElasticSearch.
  o Cross Plots, Geospatial, Time-series, Graph analysis
Geo-temporal Movement and Events View

People Trails

- By connecting users geo-coded social media posts ordered temporally aggregate “people trails” can be extracted and aggregated to see movement patterns
- User provides selection criteria to specify population segment to generate trails
  - Based upon topic or keywords, time and location or start with seed social media accounts and generate trails for community
  - E.g. generate trails for all people going through Kennedy Airport June 1 – June 20
Raw Data Drilldown
Longitude: -73.81° to -73.81°
Latitude: 40.70° to 40.70°
Total Tweets: 104
Top 10 user distribution (86 total)
Top 25 hashtag distribution (29 total)

Users
User: nytokenblackguy
Total Tweets: 5
Most Recent Tweets

September 13, 2013
NYTokenblackguy
Even as a passenger I hate the Airtrain. (© JFK AirTrain - Jamaica Station) http://t.co/u8BI1Fea57D
8:36:00 am
Geo-temporal Movement and Events View

Delta-plots and heatmaps

- Use heatmaps of activity in NYC taxi or social media to summarize overall activity and hot spots
- Delta plots show differences from average behavior or previous time periods
  - See increases or decreases in typical behavior for selected time period
- Supports identifying abnormal patterns and spatial extents
- Filtering by frequency allows focusing in on areas of largest activity or change, e.g. traffic problems during construction at the airport or change in usage during a street festival

Polar color ramps show positive / negative difference from “normal” or previous time periods

Heatmaps with a perceptually correct color ramp summarize overall activity
The delta plot data is noisy, but still a red stripe is clearly visible in this plot indicating a big change such as an event occurring.
The most frequent hashtags (TFIDF) can be used to quickly characterize the event – in this case a street festival.
6. LITIGIOUS TOPICS
Intellectual Property

- Where are the hot spots?
  ➔ What skills and expertise do we need?

- Where are the problem spots?
  ➔ More diligent in crafting patent
    (more time, more resources)
Plot showing every patent since 1982, grouped into communities based on each patent’s citation of earlier relevant patents.

© 2017 Uncharted Software
Labels, extracted using TFIDF, show the most unique terms on the most heavily cited patent in each community. There terms give an indication of the topics in each community.
And focus on this community and sub-communicates. Labels indicate topics related to telephones.
Italic angle indicates litigation ratio in that community. Steep italic indicates high litigation, reverse italic indicates low litigation.
Drill-down to communities and litigation trends.

- neuro-response data
- optically inferring jotting
- synchronization application
- object magnification
- input arrangement
- object derived object
- minimum recommendations
- caller hold queue
- tactile overlay
- portable reconfiguration
- displaying advertisements
- information retrieving
- disambiguation keypad
- recent interest
- secure transmission
- telephony computer user
- telephony platform
- illicit telephone platform
- procurement management
- interactive telephone assistant
- communicating media signals
See citations to other communities.
Method and system for preventing illicit use of a telephony platform

US 8737962 B2

ABSTRACT

A system and method for preventing illicit use of a telephony platform that includes enrolling a plurality of accounts on a telecommunications platform, wherein an account includes account configuration; at a fraud detection system of the telecommunications platform, receiving account usage data, wherein the usage data includes at least communication configuration data and billing configuration data of account configuration and further includes communication history of the plurality of accounts; calculating fraud scores of a set of fraud rules from the usage data, wherein at least a sub-set of the fraud rules include conditions of usage data patterns between at least two accounts; detecting when the fraud scores of an account satisfy a fraud threshold; and initiating an action response when a fraud score satisfies the fraud threshold.

Drill to patent publication.
6. SO WHAT?
So What?

1. Visualization isn’t just about numbers. Text repositories may be 4x or more > numeric data.

2. The next big leap in analytics and visualization will be much deeper integration of content throughout the full analytics + visualization analysis cycle.

3. Real-world applications
   - Tone analysis of your (external) communications
   - Ongoing stickiness of (marketing) messages
   - Local topics and event detection
   - Litigious communities
Text Analytics and Visualization is Much Broader

**Letters**

**Words**

**Phrases**

**Sentences**

**Paragraphs**

**Documents**

- **Categories**
- **Topics**
- **Sentiment & Emotion**
- **Social Stickiness**
- **Classifiers & Sets**
- **Graph Analysis**
- **Entity Recognition & Tagging**
- **Style Analysis**
- **Semantics/Summarization**
- **Opinion Analysis**
- **Prosody Display**

**Analysis**

- **Semantics/Summarization**
- **Opinion Analysis**
- **Prosody Display**
- **Social Stickiness**
- **Graph Analysis**
- **Entity Recognition & Tagging**
- **Style Analysis**
- **Semantics/Summarization**
What can you do?

- Understand what text data you have
- Define what business insights you want to extract
- Use open source tools for analytics and visualization
e.g. Saltlib, D3js, NLTK, Stanford Core NLP
- Incrementally experiment towards a solution
- Productionize the parts that are successful
THANK YOU

Richard Brath
rbrath@unchartedsoftware.com
@rkbrath
richardbrath.wordpress.com

Scott Langevin
slangevin@unchartedsoftware.com
@slangevi
blog.uncharted.software
About Uncharted: Profoundly visual software for improved awareness, analysis and decision-making, used by hundreds of thousands in government, law enforcement, finance, and more.