Gaining **Insight** through **Text Analytics**

Péter Molnár

Sentiment Analysis

- Process of detecting positive, negative, or neutral feelings in a piece of writing. Humans have the innate ability to determine sentiment; in a business context, however, this process is time-consuming, inconsistent, and costly.
- Steps:
  - Break the document into its basic **parts of speech** (POS) tags, which identify the structural elements of a sentence (e.g. nouns, adjectives, verbs, and adverbs).
  - identify sentiment-bearing phrases like "*terrible service*" or "*cool atmosphere*."
  - Score each sentiment-bearing phrase on a (logarithmic) scale.

Concept Matrix/Categorization

- Measure the **semantic distances between words**: semantic links.
- "Cat" is closely related to a "lion" and not an "anaconda".
- The concept of **president** can be bound to people like Obama, JFK, and Lincoln, or to positions like CEOs or Commander-and-Chief, depending on the context.
- Consider “Beverages” as a category. To build a category a **few sample words** need to be provided. E.g. *beverage*, *alcohol*, and *soda*.
- A given the sentence, “*Coca Cola returned to their original formula due to unfavourable consumer reviews*,” can be **categorized** it into "Beverages".
- Although “Coca Cola” was not used as a sample word, it was identified through the Concept Matrix with strong **semantic links** between “Coca Cola,” “beverage,” and “soda”.

Named Entity Extraction

- Automatically pulls **proper nouns** from text and determines their **sentiment** from the document. Entities like people, places, companies, brands, or job titles are classified.
- Each Named Entity has a set of associated parameters:
  - **Entity** -- The exact entity being extracted. Different names for the same reference are simplified to one name (e.g. Ol' Blue Eyes, The Chairman of the Board, The Voice, Francis Albert and Boney Baritone are condensed to "Frank Sinatra").
  - **Sentiment** -- Positive, negative or neutral tone of all mentions of an entity.
  - **Evidence** -- The number of sentiment-bearing phrases associated with a given entity.

Theme extraction

- Determine trends that appear over time. Themes are noun phrases extracted from text that can be used to identify the main ideas within your content.
- After Semantria receives the text, the engine identifies the POS tags. Two simultaneous steps occur:
- Potential themes are extracted from POS tags and kept for scoring
- **Lexical Chaining** is a process that links sentences through synonyms or related nouns to establish a conceptual chain in the content.

Summarization

- Select sentences most pertinent to the content to provide a concise synopsis of the original source text.
- Like with Theme Extraction, **Lexical Chaining** is used to select the most relevant information. The engine establishes a conceptual chain through related nouns, even when the concepts are from different parts of the document.
- The first sentence of the summary will be the longest Lexical Chain -- this should best represent the idea of your document.

Facets and Attributes

- Key points from your document and list the most important ideas with their accompanying attributes.
- **Facets** are similar to Themes, but Themes rely solely on noun phrases for analysis. Facets instead rely on Subject Verb Object (SVO) parsing, so they find trends even when there are weak or no noun phrases in your text.
- Consider the following sentence:

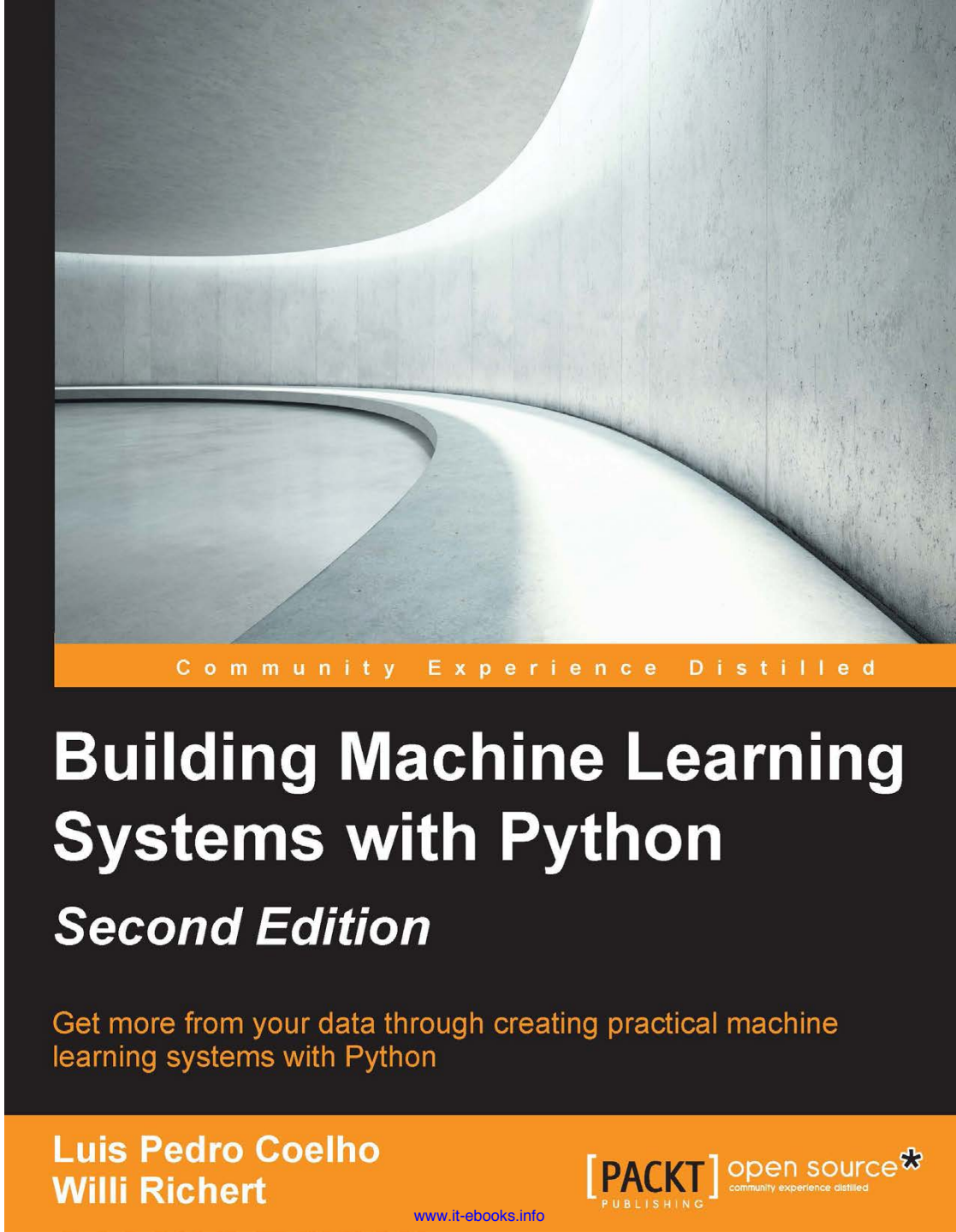
  My waiter was rude.
- Search for subject predicate object (SPO)
  In this case, "waiter" is the facet and "rude" is the attribute.

Clustering

- **Chapter 3**
  Clustering – Finding Related Posts
- Measuring relatedness, based on common words.
- Representation: Bag of words

Word	Occurrences in post 1	Occurrences in post 2
disk	1	1
format	1	1
how	1	0
hard	1	1
my	1	0
problems	0	1
to	1	0

- Representation: sparse, high-dimensional vectors
- Preprocessing: stemming, stop words, etc.



Classification

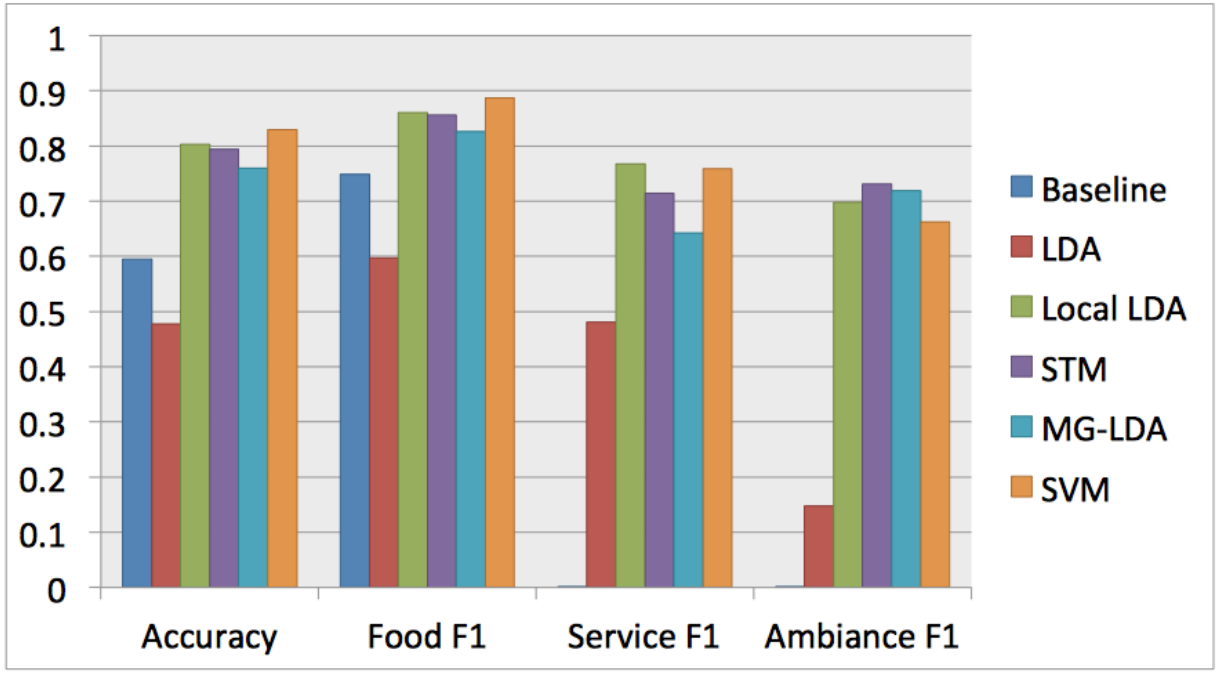
- Example:
  Movie classification by genre
- Quick read to pick-up some of the terminology.
- A few interesting links
- … ignore the XML lingo

Classification: Supervised, Labeled Data

- Need a **representation** of documents as **feature vectors**
  for example: bag of words, *(named) entities*, *sentiment*
- Need some metadata to **annotate**

Sentiment Analysis

- Different Polarities
  - Positive/negative
  - Subjective/objective
- Multiple Aspects



Multi-aspect Sentiment Analysis with Topic Models

Sentiment Analysis Issues

- Sarcasm: “It was awesome for the week that it worked.”
- Navel gazing: is when social media tracking turns up items related to your own promotional efforts, and should be filtered out.
- Neutral sentiment: is similar to the concept of swing voters
- Relative sentiment: “I bought an iPhone” is good for Apple, but not for Nokia.
- Compound or multidimensional sentiment: “I love Mad Men, but hate the misleading episode trailers.”
- Conditional sentiment: the customer isn’t angry now but says he will be if the company doesn’t call him back.
- Positive feelings can be unrelated to the core issue: many comments about actors focus on their personal lives, not their acting skills.
- Negative sentiment is not necessarily bad: Sarah Palin’s appearance on the Today show generated many negative comments but still drove ratings increases.
- Ambiguous negative words: “That backflip was so sick” is really a positive statement.

Source: http://www.adweek.com/prnewser/37705-sentiment-analysis/38709

Named Entity Recognizer

- Labels sequences of words in a text which are the names of things, such as **person** and **company** names, or locations.
- Domain specific. For smaller/specialized domains, this could be extracted manually, or from meta-data or other data bases.
- May have to depend on predefined dictionaries, such as Stanford NER http://nlp.stanford.edu/software/CRF-NER.shtml

MIDWAY WOODS
DECATUR, GA

Good Things Are Growing in the Woods

MIDWAY WOODS NEIGHBORHOOD ASSOCIATION • DECATUR, GEORGIA

Nextdoor.com Social Network

- 11,123 messages from nextdoor.com and Yahoo group
- Message and attachments
- Meta-data:
  - Sender/Neighborhood
  - Category
  - Number of replies
  - Date
- Insight:
  - Classify by category/topic, factual/opinionated
  - Predict lengthy discussion
  - Who's talking to whom?
  - Sentiment in discussions

Date: Sat, 08 Aug 2015 21:11:57 +0000
To: dr.peter.molnar@gmail.com
From: Nextdoor Midway Woods <reply@rs.email…..
Subject: Have you lost a white cat with tiger tail?
X-Received: by 10.70.129.3 with SMTP id ns3mr2909635…
Sat, 08 Aug 2015 14:11:59 -0700 (PDT)

Susan Bird from Oakhurst said:

Have you lost a white cat with tiger tail?
It has been hanging around the solarium

Shared to 13 neighborhoods
Lost & Found