Leveraging Competitive Advantage: Hidden Topics in Customer Digital Footprints

Rakesh Dhote
rakesh.dhote@gmail.com

Nov. 27, 2015

Presentation at the Scotiabank

Outline

- Executive Summary
- Topic Modeling (TM)
- Case Studies
  1. Tweet analysis – Scotiabank
  2. arXiv research articles
- Suggestions to Scotiabank

Executive Summary

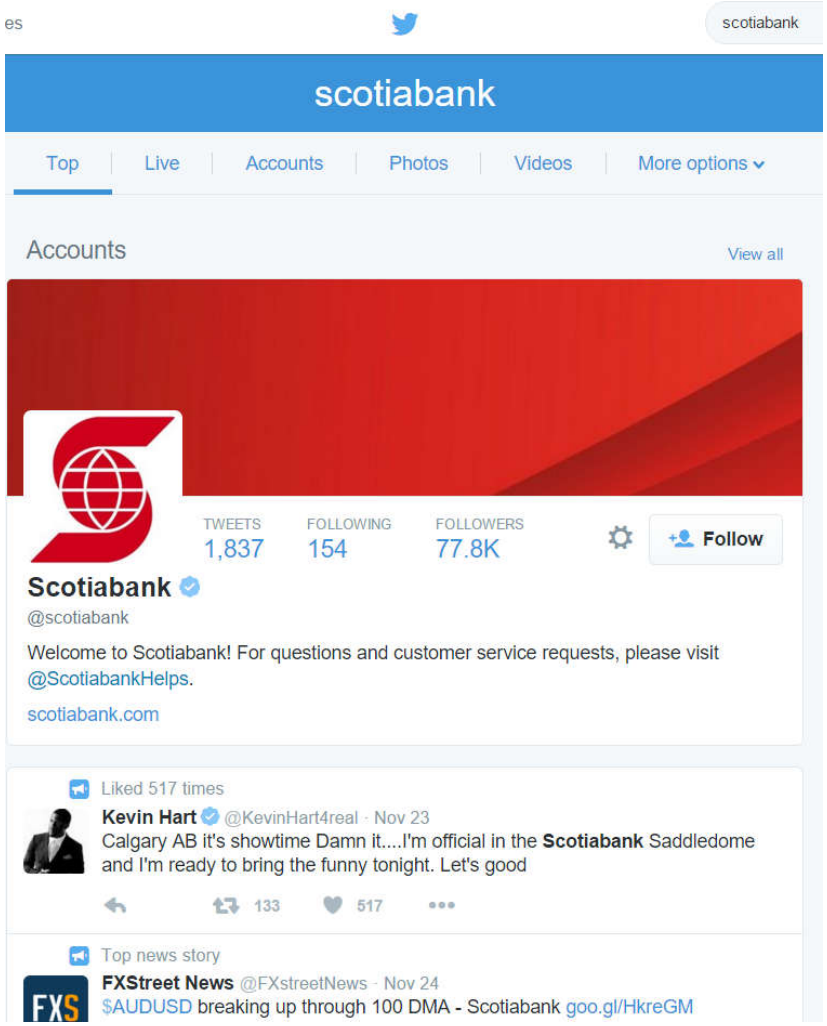
- Big Data era: customer data generated rapidly
- Social media – customer digital body language
- Actionable insights – competitive advantage
- Topic modeling
- 2 interesting case studies:
  1. Twitter data from the Scotiabank
  2. Research articles from the arXiv

Topic Modeling (TM)

- Algorithms that discover hidden (latent) topics/themes in the data
- Automatically organize, understand , search, and summarize large data
- Unstructured data – text, video, streams, etc.
- Unsupervised machine learning algorithms
  - Latent Dirichlet Allocation (LDA)
- Examples: Text Mining, Genetics, Image Tagging, Social Network, etc.

Case Studies

Case Study 1 – Tweet Analysis



- Twitter REST API
- **@Scotiabank**
- Python tweepy package
- ~**7.5k** tweets
- Data cleaning
- Latent Dirichlet Allocation
- # topics chosen = **3**
- Visualization - Wordcloud

Tweet Analysis: @Scotiabank

Topic 1

- Thank **you** Nsaks56 for following the Conspiracy of Equifax Scotiabank & LawSociety LSUC lawyers!
- Solutions Architect (Scarborough Ontario **Canada**)
- **Ticket** 3 Calgary Flames vs Dallas Stars

Topic 2

- Scotiabank is looking for a Data Scientist \ Data Engineer in Toronto apply now! **Job**
- Customer Relations Officer needed in **Toronto** at Scotiabank Apply now**!**
- **Senior** IOS Developer (Scarborough Ontario Canada)

Topic 3

- Selena Gomez concert tickets for May 17 at Scotiabank **Saddledome** in Calgary **Canada**
- Got your eye on a new oven Get it Find out **more:**
- **Calgary: Calgary Flames** vs Boston Bruins at **Scotiabank Saddledome**

Competitive (Tweet) Analysis

Scotiabank
#tweets ~ 7.5k

Topic 1

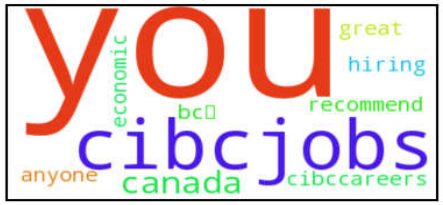
Topic 2

Topic 3

CIBC
#tweets ~ 6.4k

Topic 1

Topic 2



Topic 3

TD
#tweets ~ 5.2k

Topic 1

Topic 2

Topic 3

RBC
#tweets ~ 2k

Topic 1

Topic 2

Topic 3

Case Study 2 – arXiv articles

arXiv.org

Search or Article-id (Help | Advanced s

All papers

Open access to 1,095,499 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics

Subject search and browse: Physics Search Form Interface Catchup

30 Oct 2015: 2015 holiday scheduled announced
See cumulative "What's New" pages. Read robots beware before attempting any automated download

Physics

- Astrophysics (**astro-ph** new, recent, find)
  includes: Astrophysics of Galaxies; Cosmology and Nongalactic Astrophysics; Earth and Planetary Astrophysics; High Energy Astrophysical Phenomena; Instrumentation and Methods for Astrophysics; Solar and Stellar Astrophysics
- Condensed Matter (**cond-mat** new, recent, find)
  includes: Disordered Systems and Neural Networks; Materials Science; Mesoscale and Nanoscale Physics; Other Condensed Matter; Quantum Gases; Soft Condensed Matter; Statistical Mechanics; Strongly Correlated Elect Superconductivity
- General Relativity and Quantum Cosmology (**gr-qc** new, recent, find)
- High Energy Physics - Experiment (**hep-ex** new, recent, find)
- High Energy Physics - Lattice (**hep-lat** new, recent, find)
- High Energy Physics - Phenomenology (**hep-ph** new, recent, find)
- High Energy Physics - Theory (**hep-th** new, recent, find)
- Mathematical Physics (**math-ph** new, recent, find)
- Nonlinear Sciences (**nlin** new, recent, find)
  includes: Adaptation and Self-Organizing Systems; Cellular Automata and Lattice Gases; Chaotic Dynamics; Exactly Solvable and Integrable Systems; Pattern Formation and Solitons
- Nuclear Experiment (**nucl-ex** new, recent, find)
- Nuclear Theory (**nucl-th** new, recent, find)
- Physics (**physics** new, recent, find)
  includes: Accelerator Physics; Atmospheric and Oceanic Physics; Atomic Physics; Atomic and Molecular Clusters; Biological Physics; Chemical Physics; Classical Physics; Computational Physics; Data Analysis, Statistics an Probability; Fluid Dynamics; General Physics; Geophysics; History and Philosophy of Physics; Instrumentation and Detectors; Medical Physics; Optics; Physics Education; Physics and Society; Plasma Physics; Popular Physi Space Physics
- Quantum Physics (**quant-ph** new, recent, find)

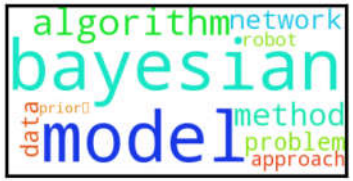
Mathematics

- Mathematics (**math** new, recent, find)

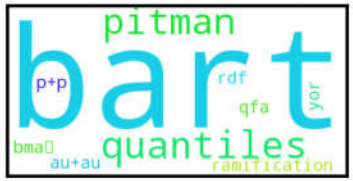
~1.1 million e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics

arXiv Topic Distributions

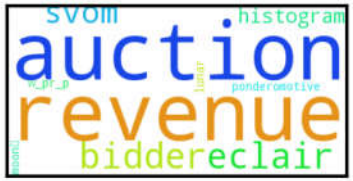
Topic 1



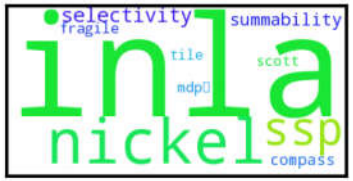
Topic 2



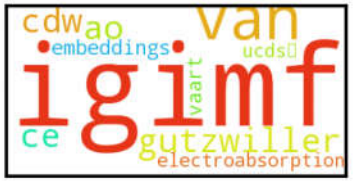
Topic 3



Topic 4



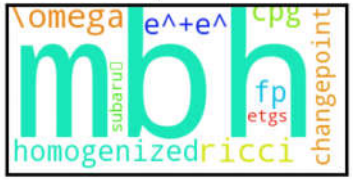
Topic 5



Topic 6

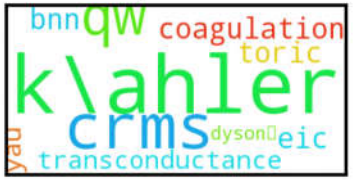
Topic 7

Topic 8



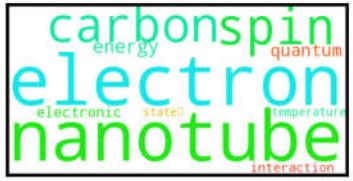
Topic 9

Topic 10

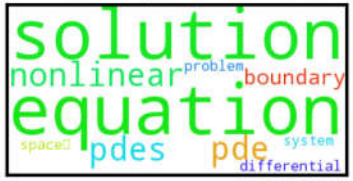


Topic 11

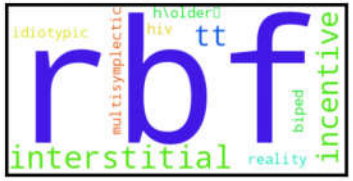
Topic 12



Topic 13



Topic 14



Topic 15

Suggestions to Scotiabank

Topic Modeling can be leveraged to find

- State-of-the-art technology in latest financial and BI research articles → promote into production
- Social media analysis → customer requirements, reduce churn rate
- Customer reviews → reveal latent subtopics
- Identify interdisciplinary field to accelerate business.

References

- D. Blei, Probabilistic Topic Models, 2003,2012
- D. Blei, Topic Models – video lectures part 1-2, 2009
- E. Chen, Introduction to Latent Dirichlet Allocation (Weblink)
- A. Oh, Topic Models Applied to Online News and Reviews (Youtube)
- Blog - Topic Modeling with Mahout on Amazon EMR
- Python Modules documentations
  - Gensim – topic modeling
  - Wordcloud – data visualization

This repository Search

Pull requests Issues Gist

rakeshdhote / Latent-Dirichlet-Allocation

Unwatch 1 Star 0 Fork 0

Code Issues 0 Pull requests 0 Wiki Pulse Graphs Settings

No description or website provided. — Edit

3 commits 1 branch 0 releases 1 contributor

Branch: master New pull request New file Find file HTTPS https://github.com/rakeshdh Download ZIP

rakeshdhote Updating ReadMe Latest commit e81ff6d a minute ago

ArXiv.txt	LDA implemenation using Python Gensim	5 minutes ago
Arxiv_test.txt	LDA implemenation using Python Gensim	5 minutes ago
LDA_Arxiv.py	LDA implemenation using Python Gensim	5 minutes ago
README.md	Updating ReadMe	a minute ago

README.md

Latent-Dirichlet-Allocation

The repo consists of the files for the Topic Modeling implemenation using the Latent Dirichlet Allocation (LDA) to uncover hidden thematic structure in the arXiv articles corpus.

The Python code uses the Gensim, NLTK, Wordcloud, and MatPlotLib packages. The program executes the following tasks:

Pre-processing:

- Cleans the arXiv articles corpus data using Python *re* package
- Tokenize, stemming, stopword removal, and lammetization (WordNet)

Thank You!