Timo Mulder

- Senior Machine Learning and NLP scientist at AMPLYFI
- MSc Artificial Intelligence, University of Edinburgh
  - Machine learning, NLP, Data Engineering, HCI
- BSc Artificial Intelligence, University of Amsterdam

Building a custom text miner … in a day

October 9, 2019

What happened?

- Needed data on a number of companies
  - What can I do in a day?
- Available: URLs to Annual reports and Interim reports per company
  - … in Excel
- **Aim: identify interesting phrases and metrics for each company that could be made ready to display in a word-cloud**

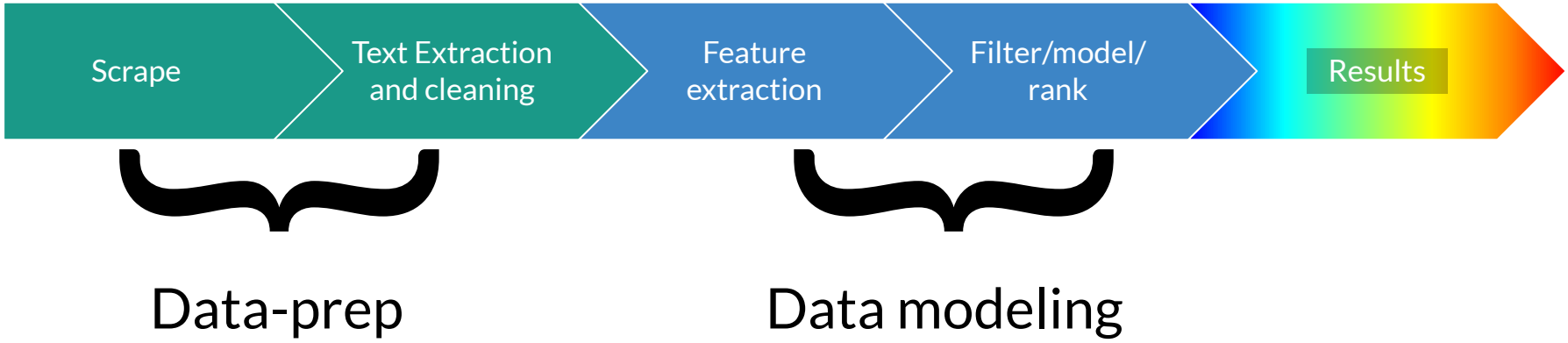
Immediate thoughts

- Simple methods
- Term Frequency - Inverse Document Frequency (TF-IDF)

Questions

- Interested in relations/discrepancies between the companies?
- How about within a company between the years?
- This deadline, why?!

Rough pipeline (and outline)



Data-prep

Scrape/download docs

- Convert xlsx to csv
- `csvreader` or `str.split(“,”)`
- Python `requests` <3
- Headers - Chromium user-agent
- Certificates, use `certifi` or simply `verify=False`

Text Extraction and Cleaning

- PDF extraction
- Paragraph deduplication for cleaning

PDF format and extraction

✅ Universal
✅ Vectors -> high resolution

❌ Vectors -> text is defined by characters and their positions
❌ Not really structured

- Pdfminer (slow and decent)
- Pdftotext (fast and decent)

Communitie.com Limited

Operating Review

Proforma operating results for the year

	Proforma 2000 Sales £000	Proforma 2000 Profit £000	Proforma 1999 Sales £000	Proforma 1999 Profit £000	Proforma 2000 Profit %	Proforma 1999 Profit %
Philatelic trading and retail operations	4,536	75	4,853	239	1.7%	4.9%
Publishing and philatelic accessories	2,557	319	2,624	392	12.5%	14.9%
Dealing in autographs, records and related memorabilia	758	243	718	226	32.1%	31.5%
Exhibitions	68	(52)	237	(22)	(76.5%)	(9.3%)
Urch Harris New Issues	-	-	512	(2)	0.0%	(0.4%)
E-commerce	98	25	-	-	25.5%	0.0%
	8,017	610	8,944	833	7.6%	9.3%
Sale of property		52		85		
Corporate overheads		(836)		(698)		
New business development costs		(413)		-		
Interest		(27)		(54)		
Before exceptional operating costs and impairment of goodwill	8,017	(614)	8,944	166	(7.7%)	1.9%
Impairment of Goodwill		(200)		-		
Exceptional operating costs		(79)		-		
(Loss)/profit before tax	8,017	(893)	8,944	166	(11.1%)	1.9%

The results above do not represent the statutory results of Communitie.com for the year as the Collector Café Group was acquired on 16 August 2000 from Flying Brands Limited. In order to present a more meaningful analysis the full year results against last year are presented which have been extracted from the audited consolidated accounts of Stanley Gibbons Holdings PLC, the holding company of the trading activities of the Communitie.com Group. Consequently, the Operating Review focuses on the proforma results for the year.

Sales

Sales for the year were £927,000 (10.4%) below last year but were only £246,000 (3%) below last year after adjusting for discontinued activities. Discontinued activities include the transfer of Urch Harris to Flying Brands Limited in June 1999 and the withdrawal from the non-profitable activity of organising exhibitions.

Sales from philatelic trading and retail operations declined in 2000 by £317,000 (6.5%). The reduction in sales is due primarily to efforts to maintain margins, reduce stock purchases and reduce debtors with the result that sales have been sacrificed to improve cash flow and bottom line profitability.

The Group lost its focus on sales during 2000 due to the absence of a detailed sales plan which is now in place for 2001. The sales plan includes a number of simple initiatives which will enhance sales driven by our recent investment in systems which enables us to analyse and target customers more effectively. Furthermore, our marketing plan incorporates specific return on investment criteria which were absent last year together with greater emphasis towards creative marketing initiatives which have been severely lacking over the past few years.

&ROOHFWRU^C&DIp^CKDV^C^T^O^U^S^S^CDUWLF
OHV^CRQ^CRYHU^C^T^S^S^CVXEMHFWV^CDQG^C^
T^O^S^S^S^COLQNV^CWR^CRWKH
U^CGHDOHUV^CDQG^CFROOHFWLEOH^CVLWHV^Q^C
^C
([SHUWLVH^CDQG^CVRIWZDUH^CGHYHORSHG^CLQ
^CRXU^CRWKHU^CVLWHV^CZLOO^CEH^CHDVLO\^C
DGDSWDEOH^CWR^C&ROOHFWRU^C
&DIp^CWR^CSURYLGH^CDQ^C
DXFWLRQ^CVHUYLFH^CDQG^CSULFH^CJXLGHV^CL
Q^CRXU^CQH[W^CSKDVH^CRI^CGHYHORSPHQW^Q^
C^C:H^CDUH^CPDNLQJ^CUHVRXU
FHV^CDYDLODEOH^CWR^CWKLV^C
SURMHFW^CWR^CVWUHWFK^CWKH^CEUDQG^CLQWR^
CRWKHU^CFROOHFWLEOH^CDUHDV^O^CDOWKRXJK^
CWKH^CPDLQ^CIRFXV^CDQG^CFR
PPLWPHQW^CVWLOO^CUHPDLQV^C
ZLWK^CWKH^CVWDPS^CEXVLQHVV^Q^C
^C
2XU^CFDSLWDO^CH[SHQGLWXUH^CSODQV^CDUH^C
VSOLW^CEHWZHHQ^CWKH^CUHIXUELVKPHQW^CRI^
CWKH^CXSSHU^CIORRUV^CDW^C
^V^\^\^C6WUDQG^O^CZKHUH^C
KDOI^C RXU^C EXLOGLQJ^C ZLOO^C

Introduction

As the newly appointed Chairman of Acme Group LLC this is my first annual report for the Group. I joined the Board in May 2016 during a very difficult period for the Group which encompassed the appointment of corporate restructuring specialists to undertake a comprehensive review of all operational aspects of the Group, a subsequent profit warning and a fundraising designed to nurse the Group through a liquidity squeeze. In short, things had gone badly adrift and urgent action was required to recover the situation. The 43% fall in net asset value, reflected in the

Introduction
As the newly appointed Chairman of Acme Group LLC this is my first annual report for the Group. I joined the Board
in May 2016 during a very difficult period for the Group which encompassed the appointment of corporate
restructuring specialists to undertake a comprehensive review of all operational aspects of the Group, a subsequent
profit warning and a fundraising designed to nurse the Group through a liquidity squeeze. In short, things had gone
badly adrift and urgent action was required to recover the situation. The 43% fall in net asset value, reflected in the

Deduplication

A systematic approach to managing risk to ensure strategic goals are met

A governance framework designed to safeguard long-term shareholder value

Read more on pages 30 and 31

2
Acme Things plc Annual Report & Accounts 2018

Read more on pages 46 to 51

Strategic report

What we do

Deduplication

Paragraph	Count
–	169
Acme Things plc Annual Report & Accounts 2016	132
29 August 2015 £m	52
Financial statements	33
3 September 2016 £m	31
For the financial year ended 3 September 2016	25
– –	22
NOTES TO THE THE FINANCIAL FINANCIALSTATEMENTS STATEMENTS CONTINUED For the financial year ended 3 September 2016	20
Governance	18
Total £m	18

Data Modeling

Feature Extraction

- Key phrase detection
- Normalisation
- Entities (such as companies)

Key phrase extraction

- Phrases tell you more than just words
  - “Store” and “cost” are individually not as meaningful as “store costs”
  - Phrases make it easy to create a narrative around them
- N-Gram extraction
  - Phrases of length N words
- Normalisation and grouping
  - e.g. stemming, removing ‘the’, ‘and’
  - To reduce sparsity
- careers@amplyfi.com

Data Modeling

- Need some sort of prevalence
- Most occuring (counts)
- TF-IDF
- Bonus - Specific Keyword Search

Order by count

Gram	Absolute_Count
financial statements	127
financial year	80
Acme Things plc Annual Report & Accounts	76
exceptional item	65
executive directors	60
Return on capital employed	56
income statement	51
Audit Committee	40
external auditor	39

TF-IDF

- **Term Frequency - Inverse Document Frequency**
- Term frequency = total in current document
  - Higher frequency -> importance -> higher value
- Document frequency = # of docs term appears in
  - Lower frequency -> uniqueness -> higher value

$$w_{t,d} = (1 + \log(tf_{t,d})) \log(\frac{N}{df_t})$$

Choose your corpus wisely

- **One year, all companies**
  - All companies talk about term -> low score
    - e.g. “online competitors”
  - Company specific terms -> high score
    - e.g. “Acme Things Plc annual report”
- **One company, all years**
  - Retrieve key events and focus per year

Bonus - Chosen Keywords

- Keywords such as acquisition, growth, risk
- Percentage of all words
- Keywords in key phrases
- Stemming, acquisition -> acqui

Results

Results - old

Gram	Weight	Absolute_Count
financial statements	0	127
financial year	0	80
Acme Things plc Annual Report & Accounts	0.967	76
exceptional item	0.449	65
executive directors	0	60
Return on capital employed	1.41	56
income statement	0	51
Audit Committee	0	40
external auditor	0	39

Results - old

Gram	Weight	Absolute_Count
financial statements	0	127
financial year	0	80
Acme Things plc Annual Report & Accounts	0.967	76
exceptional item	0.449	65
executive directors	0	60
Return on capital employed	1.41	56
income statement	0	51
Audit Committee	0	40
external auditor	0	39

Results - new

Gram	Weight	Absolute_Count
Loss on disposal and write-off	2.274	11
Ernst & Young LLP	2.055	7
Jane Doe	2.055	7
onerous lease charges	1.981	6
onerous lease provision	1.981	6
Disclosure Guidance and Transparency Rule	1.893	5
Free cash flow	1.893	5
AIM	1.893	5

Keyword search results

keyword	count	percentage
risk	161	0.185
growth	96	0.11
acquisition	18	0.021
acquired	9	0.01
acquire	4	0.005

Gram	Count	Percentage
principal risks	10	0.164
risk of fraud	5	0.082
principal risks and uncertainties	5	0.082
growth rates	5	0.082
Foreign exchange risk	5	0.082
risk management	4	0.066
risk appetite	4	0.066

- “Digital growth”

Further steps

- Find context
  - Relation extraction
  - Term co-occurrence
- Additional filtering
- Topic modeling (clusters of words or phrases)
  - K-means clustering on vectors - simplest
  - Latent Dirichlet Allocation (LDA) - probabilistic, most famous
  - Matrix Factorisation (dimensionality reduction)

Conclusions

- Python is ideal for quick scraping of documents
- PDF is a painful format for text extraction
- N-Grams say more than words
- TF-IDF - simple, quick, quite powerful
- Further steps possible

Questions?

Further Reading

Edinburgh Natural Language Understanding course:

http://bit.ly/uoelanguage