Text mining to correct missing CRM information

A practical data science project

Python Ireland 09 April 2014
Jonathan Sedar, Applied AI
@jonsedar

In a nutshell

- A CRM dataset (100k business accounts) belonging to a national energy supplier
- A knotty problem: multiple accounts per company, without any grouping ids
- How can we to find groups of accounts (larger company structures), using just the CRM data?
- Machine Learning (ML) and Natural Language Processing (NLP) tools and techniques in Python.
- Import: Scikit Learn and TextBlob (NLTK & Pattern)

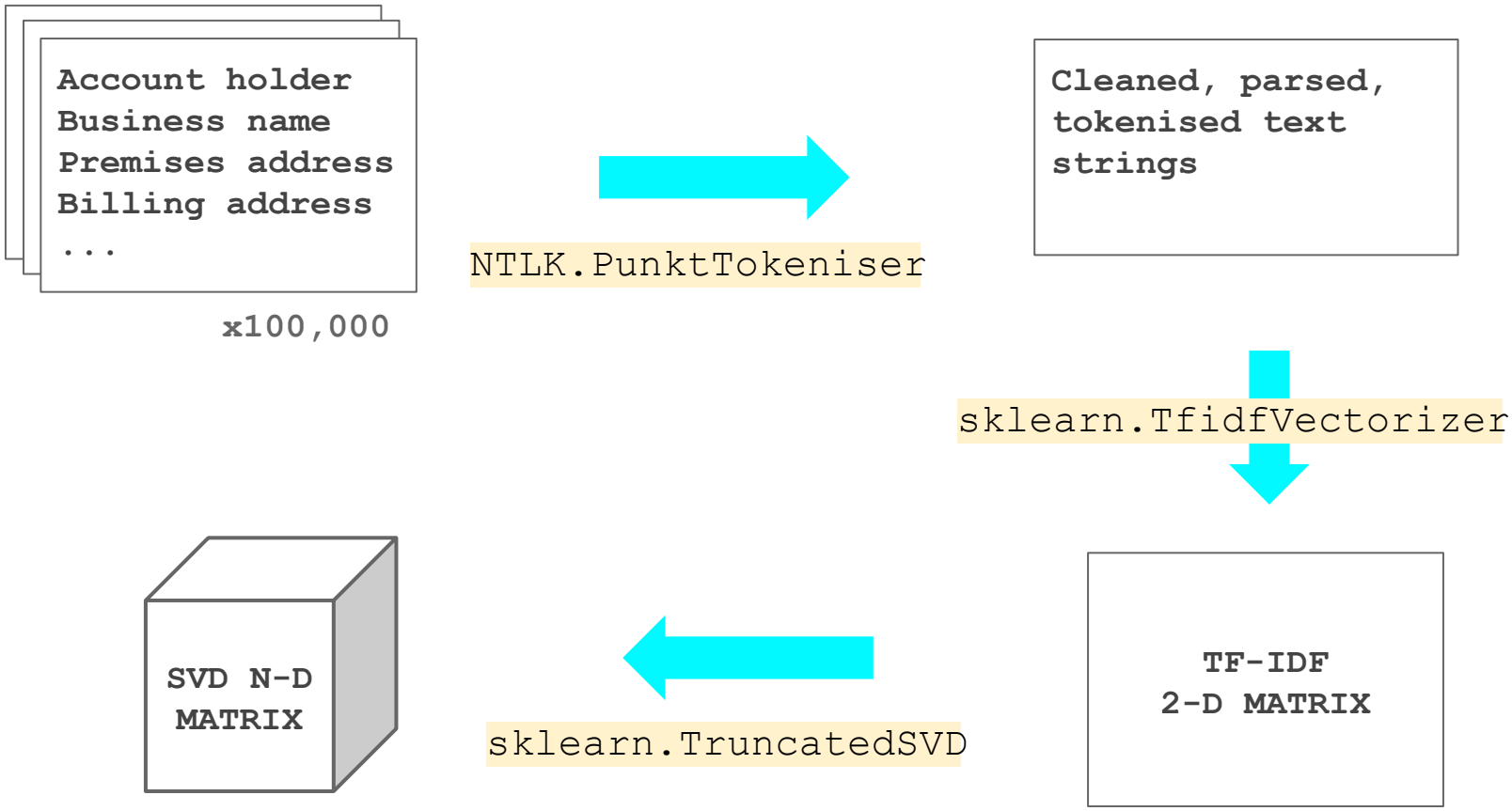
Show me the data!

Company ID	Account name	Contact name	Premises address lines 1 - 4	Billing address lines 1 - 4
1	Bob's Pizza	Big Bob	5 High St, Wexford	5 High St, Wexford
1	Bob's Pizza	Big Bob	Temple Bar, D2	5 High St, Wexford
1	Mike's Kebabs	Mad Mike	3 Upper St, Dublin	5 High St, Wexford
2	Mark's Kebabs	Mild Mark	8 Upper St, Dublin	Main St, Waterford
3	Fred's Falafel	Fat Fred	9 Henry St, Cork	9 Henry st, cork
3	Fred Fallafell	Freddie	Bridges St, Galway	Henrys St, Cork

This crucial bit of info groups the separately-recorded accounts into companies…
and was missing from the dataset

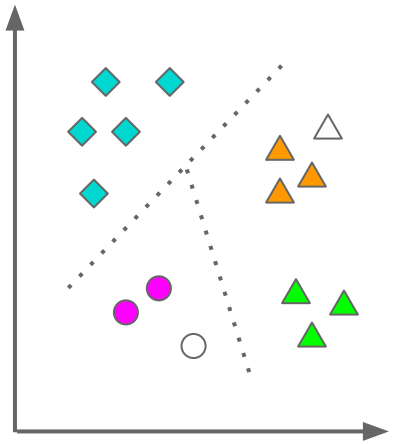
… x100,000

Transforming text into useful structures



Now we can identify similar accounts:

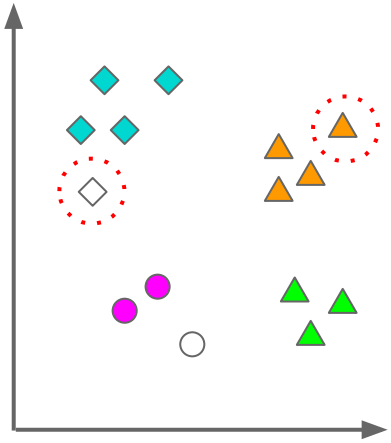
1. Suggest similar accounts to be grouped



`sklearn.MiniBatchKMeans`
`sklear.AffinityPropagation`

2. Human validation & verification

3. Incorporate & propagate valid groupings



`sklearn.RadiusNeighborsClassifier`

Show me

- Created an IPython notebook to demo the principles using an analogous dataset
- Code hosted on github at

  `https://github.com/jonsedar/textmining`

A pragmatic process using OOTB Python machine learning and human expertise

- A very quick turnaround from raw data to tagged companies to 93% accuracy
- ~40% of accounts found to belong to a company, ~3.5 accounts per company
- NLP toolkits and scikit-learn allowed rapid development and testing of solution
- Incorporated human identification at critical stages: no ML problem is an island

Thank you

Any questions?