Data Science Vignettes in Java

Rui Miguel Forte
Lead Data Scientist @ Workable

Data Science

This looks cool, but what is it really?

Data Science

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

Data Science

- It is a fertile mixture of:
- Using statistical and predictive modeling techniques on data in order to build models that describe and/or predict some process
- Implementing, evaluating and maintaining (testing?) these models using software
- Delivering a result that is somehow useful to a business

Content Analysis with Apache Tika

Getting the most out of your documents

Our Problem:

We wanted to be to automatically to perform text mining on different documents corresponding to uploaded candidate resumés

Apache Tika

- Apache Tika (http://tika.apache.org) is a content analysis toolkit
- It powers text extraction portion of projects such as Apache Solr
- It has functionality for:
- Detecting document types, encodings, languages
- Extracting metadata, embedded links and images

Apache Tika

- Documentation on the website is excellent
- Some of the more advanced functionality such as nested image extraction is hard to find
- Also, the Manning book is really old now
- In order to work with different types of documents it uses projects such as PDFBox
- It has a quite frequent release cycle (~6 months) and at the time of this, it is in version 1.13.

How we Use Apache Tika

- Apache Tika enables us to extract information from candidate resumés in real time as soon as they are uploaded
- This information is then fed into a custom Resumé Parser that we have written in Java

Demo Time!

Do try this at home folks and suggestions (a.k.a. pull requests) very welcome!

https://github.com/ruimiguelforte/tika-tutorial

I am using git-flow so you will find the current code on the develop branch.

Text Extraction Issues

- Header/Footer info may appear interspersed in the text e.g. page numbers
- Layout/markup information is lost
- We can retain some layout information for MS Word documents
- Characters may be altered (accents), repeated (strangely enough), spaced out (due to weird fonts or markup), lost (due to encoding issues)

Text Extraction Issues

- Information may appear out of order due to columns, tables, word art etc…
- Some warnings may appear even if extraction proceeds regularly.

Porting Machine Learning Models

Working with PMML

Our Problem:

Often, we train a model using an environment such as R or Python but want to deploy in Java for production

The PMML Format

- The PMML format is an XML based format developed by the data mining group (www.dmg.org) whose goal is to define a standard way for representing and exchanging predictive models.
- The group is working on a JSON based format but it is still under development

The PMML Format

- Over the years, the PMML format has increased its coverage of the different models out there including random forests, neural networks etc…
- Many environments such as R, Python and Java have packages available for importing and exporting models using this standard

How we Use PMML

- In Java we use jpmml-evaluator (https://github.com/jpmml)
- This allows us to import models that we train in R or scikit-learn and deploy them in Java
- One important note to make is that model features should ultimately be computed in one place
- The code for features used during model training and model prediction should be the same

Demo Time!

Do try this at home folks and suggestions (a.k.a. pull requests) very welcome!

https://github.com/ruimiguelforte/jpmml-tutorial

I am using git-flow so you will find the current code on the develop branch.

Thank you

Rui Miguel Forte
miguel@workable.com
Data Science Athens Meetup http://bit.ly/1OQPVwU
My book: http://bit.ly/1iDIIY1