Information extraction from business documents with machine learning
<table>
<thead>
<tr>
<th>Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data and Text Mining</td>
<td>3</td>
</tr>
<tr>
<td>CV parsing project</td>
<td>7</td>
</tr>
<tr>
<td>Programming approach</td>
<td>11</td>
</tr>
</tbody>
</table>
**Data and Text Mining**

Some definitions

- **Data mining** is the computing process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and database system.

- **Text Mining** is a special form of **Data Mining**, applied to “unstructured” texts (press agencies, web pages, e-mails, etc.) and more generally to any document corpus.
A Text Mining process is generally structured in four phases:

1. Data acquisition
2. Preprocessing
3. Modeling
4. Results validation
Text Mining

Text Mining: preprocessing

- In the **preprocessing** phase, the linguistic analysis is performed and all that is needed to arrive at a **vector representation** of the document is done. In particular:
  1. POS tagging
  2. Lemmatization/Stemming
  3. Definition of stop-words
  4. Dimensionality reduction
  5. Meta-information integration
  6. ...
In the **modeling** phase, the vectorized documents are subject to a **machine learning algorithm**, specific to the target.
CV parsing project

Overview

Description

• **Information extraction**
  – Automatically analyze business documents as they flow through business communication channels
  – Extract information from **Italian** unstructured documents

• **Textual data**
  – Test: CV
  – To Be: enterprise documents for compliance check
We use a combination of unsupervised and supervised methods to extract information from Italian unstructured documents.

Combination of **unsupervised** and **supervised** classifiers to decide whether a piece of text represent a certain information or not.

**Information classes**
- Personal information
- Skills
- Education
- Work experience
Classification Method

Machine Learning Algorithms: details

- We apply a three step classification, where methods in the previous step creates the features for the classifiers of the following step.

1st step

- **Word embeddings**
  - Word2vec

- **Named Entity Recognition**
  - Stanford NER tool

- **Custom task features**

2nd step

- **Supervised neural network classifier**

3rd step

- **Information class association**
  - (e.g. threshold classifier)

- For every word is calculated the **probability to represent a certain information** (e.g. name, surname, skill,...)

- For each information class we chose the **words with high probability**

- Every piece of **text** is tagged with **different methods**

- Every word is enriched with the information extracted by each method
Final Product

Business Document Analyzer

Machine Learning Kernel

Frontend
Programming approach
Research and Industry approaches

**Research**
- **Ease** of implementation
- **Testing** many different models
- **Availability** of scientific libraries
- **Flexibility**

**Production**
- **Stability** reliably running for long time
- **Scalability** to different amount of data
- **Robustness** to a range of different conditions
- **Integration** with the company infrastructure
From R&D to production code

Challenge of Building a Product

• Writing production code:
  1. **Engineering** the product for expandability and maintainability
  2. Chose the right **tools** and **programming** languages
  3. **Optimize** (memory footprint and execution speed)
  4. Adapt to **specific hardware** (e.g. GPU, Clusters)
  5. Adapt to **specific libraries** (e.g. distributed computing libraries)
  6. **Integrate** the entire product pipeline (data managing, web interface, integration with third-parties services)
Conclusion

The mission

- Finished products
- Companies interests
- Investments for applied research projects
- Publications & Prototypes
- Software Engineering

The mission.
THANK YOU
FOR YOUR ATTENTION

Niccolò Fava
Software Developer

niccolo.fava@res-it.com