Das ist MindCraft.ai

Decision-making Engines for Data-driven Businesses, especially:

- Document and Web pages Classification, Capturing (NLP, CNN, CV, NER)
- Statistical Prediction (DNN, Regression, Prognosis)
- Command Centers for IoT systems (RNN, Time Series, Anomaly Detection)
- Data Analysis and BI
Document classification:
- Document processing
- Web scraping
- GDPR and other laws
- Poor quality
- Handwritten text
- Different templates
NER or named entities recognition:

- OCR engines
- Document formats
- General or document specific (e.g. date)
- Columns and other formatting
- Abbreviations
Typical NLP process:

- OCR
- Cleaning (symbols, stop-words, etc)
- Tokenization (sentence and word)
- Embeddings (BOW or word2vec)
Issues with the typical process:

- Missed information (color, style, template)
- Tokenization of German-like languages
- Arabic and Hebrew languages
- Table columns
- Contextual labeling is expensive
polyglot:

- 40 models
- Trained on Wiki
- Semi-supervised learning
- Embeddings plus Shallow Neural Network

<table>
<thead>
<tr>
<th>Language</th>
<th>Sentence</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Simlen was traded from the Heat along with Antoine Walker and Michael Doleac to the Minnesota Timberwolves on October 24, 2007, for Ricky Davis and Mark Blount.</td>
<td>-</td>
</tr>
<tr>
<td>Hungarian</td>
<td>Dimitri beszélt egy utat Rómába.</td>
<td>Dimitri talked about a trip to Rome.</td>
</tr>
<tr>
<td>Spanish</td>
<td>Pešek nació en Praga y estudió dirección de orquesta piano en la Academia de Artes allí, con Václav Smetacek.</td>
<td>Pešek born in Prague and studied orchestra direction piano at the Academy of Arts there, Vaclav Smetacek.</td>
</tr>
<tr>
<td>Russian</td>
<td>Уроженец Рио-де-Жанейро, Хосе Родригес Тринда-дади использовал сокращенную форму как своим специфическим псевдонимом.</td>
<td>A native of Rio de Janeiro, Jose Rodriguez Trindade used as a shortened form of his stage name.</td>
</tr>
<tr>
<td>French</td>
<td>La France veut satisfaire à ses engagements envers l’Union européenne.</td>
<td>France wants to meet its commitments to the European Union.</td>
</tr>
<tr>
<td>Turkish</td>
<td>Erdoğan, Türkiye’de twitter yasaklandı.</td>
<td>Erdogan banned twitter in Turkey.</td>
</tr>
<tr>
<td>Arabic</td>
<td>قال غاري بيل، حكم فريق ريال مدريد، إن الت流入 يدوري أبطال أوروبا سيبطل معه إلى الأبد. ولذلك بعدما ساعد هذه الائتماني الفاكرية فريق للنادي الملكي على هزيمة نادي اللينوكس مقردرين في Spain.</td>
<td>Said Gareth Bell, the star of Real Madrid, said that winning the Champions League will remain with him &quot;forever&quot;, and that after his goal helped the club to defeat Atletico Madrid in Spain.</td>
</tr>
<tr>
<td>Indonesian</td>
<td>Rendjambé meninggal dalam keadaan tidak jelas , yang mengakibatkan kemewahan oleh pendukung oposisi marah di Port Gentil - dan Libreville.</td>
<td>Rendjambé died in unclear circumstances, which resulted in riots by angry opposition supporters in Port Gentil - and Libreville.</td>
</tr>
<tr>
<td>Chinese</td>
<td>新华社 25日报道称，新疆公安局。</td>
<td>Xinhua News Agency report quoted the 25th Xinjiang Public Security Department.</td>
</tr>
<tr>
<td>Greek</td>
<td>Ενθάμε βγήκε τα Γκούτιονιν κάτω από το λόγμα ανατολικά του Τύγρη, να φέρει το θάνατο σε ελλάδικη τη Μεσοποταμία.</td>
<td>Ennil brought the Gutiens down from the hills east of the Tigris, to bring death throughout Mesopotamia.</td>
</tr>
</tbody>
</table>
spaCy:

- Subword embeddings
- Affixes
- Convolutions with residuals
- No Bi-LSTM
- Custom models
- Active Learning
Ready-Solution problems:

- Missed capital letters info
- Disambiguations (Bank or Bank, numbers)
- Work for generic terms, not for document specific
- Only contextual labels
- Only basic ENs (PER, ORG, LOC)
- Only few languages working good
How we do it - Document classification:

- Typical BOW model
- CV analysis using CNN or Shallow NN
- $70 + 70 = 93$
How we do it - Location model:

- statistical model for every EN location on the doc
- 62% accuracy out of 22 docs
How we do it - Word model:

- word2vec embeddings
- Horizontal and vertical tokenizer
- 1D CNN + Bi-directional LSTM
- 80% accuracy out of 12 ENs
How we do it - Character model:

- Feature engineering (style, capitalization, color, punctuations)
- Bi-directional LSTM
- Hysteresis
- Cheap automatic labeling
- Data Anonymization
How we do it - stacking models:

- RE filters
- Joining contexts (street and number)
- Entity linking (removing disambiguation)
- Stacking horizontally as features
- Use Bayesian rule
- Test accuracy over 95%
30 years ago: ML model to play checkers (thanks to my father and Martin Gardner) (picture from www.instructables.com)

10 years ago: ML framework in .NET that has tf-like operators, optimizer and even automatic hyper-param tuner
Das ist MindCraft.ai

Decision-making Engines for Data-driven Businesses, especially:

- Document and Web pages Classification, Capturing (NLP, CNN, CV, NER)
- Statistical Prediction (DNN, Regression, Prognosis)
- Command Centers for IoT systems (RNN, Time Series, Anomaly Detection)
- Data Analysis and BI