Data Science at DENIC

Our exploits in the realm of Analytics

Boban Ben Dodevski (Business Data Analyst)

APTLD74, Tashkent, 21.09.2018
AGENDA

1. Personal Introduction

2. Planned Projects at DENIC with an analytics focus
   1. Central BI Dashboard: DENIC Domain Intelligence
   2. The Crawler Project
   3. DNS Analytics

3. Contact

4. Links
1. Personal Introduction

• Business Data Analyst at DENIC since 04/2017
• Background in Economics and Statistics
• My duties can be separated into two categories:
  • On Business side
    • Observation of market trends that are changing the (domain) Industry
    • New developments like Blockchain (DNS), IOT, Security topics with an analytics focus (Anomaly Detection)
    • What are other registries doing (collaboration i.e. Registry Data Nerds) and checking where we stand
  • On Analytics side
    • Analysing our NIC Database (good part of 2017)
    • Ad-Hoc Analytics tasks
    • Pushing forward the field of Analytics at DENIC
AGENDA

1. Personal Introduction

2. Planned Projects at DENIC with an analytics focus
   1. Central BI Dashboard: DENIC Domain Intelligence
   2. The Crawler Project
   3. DNS Analytics

3. Contact

4. Links
2. Planned Projects at DENIC

• Key goal of every analysis is to generate insights from analysing Data on which we can act upon
• To achieve this goal there are currently three projects on our roadmap:
  a. The DENIC BI Dashboard
  b. The Crawler Project
  c. DNS Analytics
• Due to more important projects they are all still in an conceptional stage
• Visuals for the next time I present to you
AGENDA

1. Personal Introduction
2. Planned Projects at DENIC with an analytics focus
   1. Central BI Dashboard: DENIC Domain Intelligence
   2. The Crawler Project
   3. DNS Analytics
3. Contact
4. Links
2. Planned Projects at DENIC

1. DENIC Domain Intelligence (DDI): The central BI Dashboard

What is the goal of the project?
- The purpose of this project is build a platform on which to present data (with a more economic character)

Who will be using it?
- Members can keep track of recent developments (on a glance)
- Our Management can monitor important metrics on a regular basis
- Small public dashboard for the interested

What kind of data will be presented in the Dashboard?
- NIC, our active registry database (Registrar, Registrant)
- Crawler Data (see 3)
2. Planned Projects at DENIC

1. Content: What will be on the Dashboard?

- Because we don’t have a finished dashboard yet I can only show what I planned and not the final result
- Members want to compare themselves with each other to see how they perform -> Performance measures
- The Dashboard will have four main sections:
  a. Domain Movement
  b. String Analysis
  c. Summary Statistics
  d. (Fun Facts)
2. Planned Projects at DENIC

1. DDI Mindmap
2. Planned Projects at DENIC

a. Domain Movement: Bread and Butter Metrics

- Time Series of movements per member (CREATE, DELETE, …)
- Time Series of movements per Business Model (absolute values will be masked as rate in order to satisfy member confidence)
  - Performance measures for our Domain Investors
- Will be updated daily
2. Planned Projects at DENIC

b. String Analysis (Text Mining)

- A domain is a string and therefore can be analysed using Text Mining
- Discovery of interesting facts, some of them relevant for security
  - Amount/Distribution of Domains starting with a specific character (a-z, 0-9)
  - Composition of a domain (characters, numbers, dash)
  - Character Sets (Latin, Arabic, Chinese, IDNs, …)
  - Top Strings [lawyer-xy.de, first-lastname.de]
    - -> Useful for automatic Domain suggestion
  - Abuse investigation [amazon.de, buy-amazon.de, amazontravel.de]
    - -> purely informative reasons (research): We don’t act on Content!!!
  - People/Bots can create Domains, Domain Generating Algorithm (DGA), Random domain generation causes domains to have a high level of entropy [denic.de, dahg7z-k567o-pyw.de]
    - -> purely informative reasons (research): We don’t act on Content!!!
2. Planned Projects at DENIC

c. Summary Statistics

- Members per country/region or Domains registered per administrative unit
  - Regional Statistics which you can find on our website
- Domain Growth excluding certain high performing members
- Deletion rate, Creation rate and other indicators
- Owner of websites (Companies Type, Privateers, Organizations)
- Drop catching Rate of Domains
- And many more …
AGENDA

1. Personal Introduction
2. Planned Projects at DENIC with an analytics focus
   1. Central BI Dashboard: DENIC Domain Intelligence
   2. The Crawler Project
   3. DNS Analytics
3. Contact
4. Links
2. Planned Projects at DENIC

2. The Crawler Project (Joint project of the Registry Data Nerds)

What is the goal of the Project?
- A crawler (or web fetcher) is a piece of software that gathers data
- Generating insights by collecting additional data (enriching, important in an stagnating/receding domain market)
- Quite a challenging project when crawling a large amount of websites

Who will be using the results?
- Members > possible upselling opportunities
- Management > Better understanding of industry, new revenue sources

What kind of data will be used?
- A scan of all domains in our Zone-file, extraction of 30+ metrics like html content of a page, check of certificate validity, geolocation of server, DNS Route taken to final page, SSL certificate Information, …
2. Planned Projects at DENIC

2. Primary Question of Interest and how to find the answer

• How can we classify all website(s) from our Zone into a fixed set of categories?

• This is a classical machine Learning Problem (Multiclass, Single Label, Hard Classification) and we need three things for that:
  1. Content of a website (html header, body, and other stuff)
  2. A clearly defined set of Categories -> The registry Data Nerds developed one called Domain Industry Taxonomy (DIT)
  3. A training set where a certain amount of websites where classified into a category
  4. After (a lot) of Data Cleaning the Algorithm starts to work
  5. Result can be considered a Map of the .de Zone
2. Planned Projects at DENIC

2. Classification Problem: Example

- Determining the share of our members in a certain Industry i.e.
  - 100,000 Engineering companies in DIT 1002 and Member 002 has 10,000 of them under contract -> 10% Market share of all domains
- Further enriching this with Data the amount of companies in a certain industry in a country will tell you the overall market penetration i.e.
  - 12,000,000 lawyers in Germany but only 12 of them have a website ->

It might be worth going to the Annual Mega Law Symposium for PR
AGENDA

1. Personal Introduction
2. Planned Projects at DENIC with an analytics focus
   1. Central BI Dashboard: DENIC Domain Intelligence
   2. The Crawler Project
3. DNS Analytics
3. Contact
4. Links
2. Planned Projects at DENIC

3. DNS analytics

- We are in the early stage of evaluation so not many details yet
- Operational Relevance
- Information Security
- Forensics
- Developing Security related data products for internal use
4. Contact

Boban Ben Dodevski
Business Data Analyst
dodevski@denic.de
5. Links

DIT  Domain Industry Taxonomy on CENTR
DGA  Domain Generating Algorithm
Classification  Web Page Classification: Features and Algorithms
Classification 2  Classification of Websites by Industry Sector