Subjects clustering

Jakub Vonšovský
Brno 2015

Motivation

- I completed graphical design and typography – this area interests me so what next?
- Recommender system for IS offering subjects similar to those I completed in development (Hana Bydžovská)
- Each cluster contains subjects similar to each other

Introduction

- 3-phase algorithm in which data preparation, distance measurement and making clusters takes place
- First two phases in Python language (own work), together about 45 KB
- Clustering algorithm is external application in Java language with author's persmission, from my side programming of common interface, cluster measurement metrics and understandable (for our purposes) output

Phase I – Data grabbing

- First version used given csv files – small set of subjects (63) and also not current
- Currently when list of subjects is changed (now all data about all subjects at faculty of informatics are included) information about literature, anotation, … should be downloaded
- For time reasons downloading of grades was removed / not completed as required url is different from semester to semester, faculty to faculty

Phase I – Parsing demo

```
# dvě a více jmen
m = re.search('^(['+word+'\s\.\-]+?)\s*[\\\&,]+\s*(['+word+'\s\.\-]+?)(\s*,\s*['+word+'\s\.\-]+)?[:\.]', record)

if m:
  authors.append(m.group(1).strip().replace(".", ""))
  authors.append(m.group(2).strip().replace(".", ""))
  if m.group(3):
    authors.append(m.group(3).strip().replace(".", ""))

else:
  # Edited by
  m = re.search('Edited by (['+word+'\-]+\s\w\.\s['+word+'\-]+|['+word+'\s\-]+)(\s\-\s)?' +
    '(['+word+'\-]+\s\w\.\s['+word+'\-]+|['+word+'\s\-]+)?\.', record)
```

Phase II – Distance matrix

- For each attribute – teachers, fields, prerequisites, field constraints, supervisors, literature, objectives and syllabus (and grades) distance of two subjects is calculated
- These parameters are then weighted and sum in just one number
- As bigger number means the better in this case it has to be converted then to distance values where smaller means the better

Phase II – Normalization

- For each row in matrix:

- First method uses similarity to itself as base number and all values are subtracted from it
- Second method tried to enhance first one by normalize these values into <0 + ε, 1>
- Third method uses logistic function $P(t)=\frac{1}{1+e^{-t}}$

  to avoid getting close to 0 or 1

Calculating distance – Jaccard Index

- $J(A,B)=\frac{A\cap B}{A\cup B}$
- Simple metric used for most of attributes (authors of literature, fields, …)
- For teachers values of set have also weights – lecturer, deputy have 1, seminar tutor 0.5 and assistants have no value as they probably don't have relationship with subject area itself and fluctuate
- At first also for text but then TF-IDF was added

Calculating distance – TF-IDF

- TF-IDF of word in document is counted by number of its occurences divided by number computed by occurences of this word in all documents (in this context document is text from syllabus or objectives)
- This value is computed for every word in every document (even if it's zero)
- Then cosine similarity is launched on there vectors giving again only one similarity number

Calculating distance – text generally

- How to store words in czech fusional language?
  - First simplistic approach uses only words longer than 3 letters and stores maximum of 6 letters
  - Second, "smarter" approach uses stemming application I wrote several years ago
- After few tests it showed that TF-IDF is better than Jaccard and storing 6 letters is better than stemming (am I bad / dumb programmer?)
- However TF-IDF has high complexity of $O(s^2*w)$ (subjects, words) running then 20 minutes instead of 1,5 making it bad bottleneck

Calculating distance - prerequisites

- Problematic part of application – for example it sometimes produces higher similarity number for some other subject then to subject itself
- When one subject is prerequisite of some other subject, it's taking advantage by high weight into final sum
- Same for "banned" subjects – they have to be punished (by high weight)
- Subjects with no relationship have prerequisite weight set to 0

Normalization results

- After manual evaluation it seems that first method of "reverse values" is again better than other two "smarter" approaches
- On testing data the other two methods for example put "Artificial Intelligence I" next to "Essentials of General Logic", another example on full data is "Bioinformatics I" and "Computers and Ergonomy"
- Maybe sparsely distributed values are better than dense as it is done artificially

Phase III - Clustering

- After two phases application launches clustering algorithm based on Murtagh Average Clustering Linkage where the only input parameter is maximum cluster size (threshold)
- Davies-Bouldin and Dunn index were implemented to measure "quality" but it showed that it is not very usable in our context – these two measures went against each other many times, "good" numbers were for not very usable clusters, …

Results

```
Cluster size: 10
PV171;Diagnostika číslicových systémů
PV172;Architektura digitálních systémů
PV198;Aplikace jednočipových počítačů
PB170;Seminář z konstrukce digitálních systémů
PA174;Design of Digital Systems II
PV170;Konstrukce digitálních systémů
PB171;Seminář z architektury digitálních systémů
PA175;Digital Systems Diagnostics II
PA176;Architecture of Digital Systems II
PV191;Projekt z konstrukce digitálních systémů
```

```
Cluster size: 10
SDIPR;Diplomová práce
SBAPR;Bakalářská práce
SZMGR;Státní zkouška (magisterský studijní program)
SZBIO;Státní zkouška (bakalářský studijní program dvouoborový, Informatika)
SZB2;Státní zkouška (bakalářský studijní program)
SZB;Státní zkouška (bakalářský studijní program)
SZB1;Státní zkouška (bakalářský studijní program)
SZMIO;Státní zkouška (magisterský studijní program, Učitelství pro SŠ)
SZB3;Státní zkouška (bakalářský studijní program)
SOBHA;Obhajoba závěrečné práce
```

```
Cluster size: 8
PA182;Managing in Reality
PV237;Strategy and Leadership
PA180;Interim Project Business
PV236;Time Management and Effectiveness
PA186;Interim Project - Research II
PA185;Interim Project - Research I
J004;Intercultural Management
PA194;Introduction to Service Science
```

```
Cluster size: 6
PV131;Digitální zpracování obrazu
PB130;Úvod do digitálního zpracování obrazu
PV162;Projekt z digitálního zpracování obrazů
PA173;Mathematical Morphology
PV121;Počítače a hudba I
PV129;Počítače a hudba II
```

Three methods

Reverse values

Cluster size: 4
IA080;Seminar on Knowledge Discovery
PV115;Laboratoř dobývání znalostí
PV056;Strojové učení a dobývání znalostí
PA164;Strojové učení a přirozený jazyk

Cluster size: 4
IB031;Úvod do strojového učení
IA158;Real Time Systems
PV021;Neuronové sítě
IV125;Seminář laboratoře Formela

Cluster size: 1
IB101;Úvod do logiky

Cluster size: 1
IA008;Computational Logic

Reverse values to <0 + ε, 1>

Cluster size: 5
IA080;Seminar on Knowledge Discovery
PV115;Laboratoř dobývání znalostí
PV056;Strojové učení a dobývání znalostí
IA008;Computational Logic
PA164;Strojové učení a přirozený jazyk

Cluster size: 5
IV003;Algoritmy a datové struktury II
IB110;Základy informatiky
IB101;Úvod do logiky
IB000;Matematické základy informatiky
IB002;Algoritmy a datové struktury I

Cluster size: 4
IB031;Úvod do strojového učení
IA168;Algorithmic game theory (slovo Bayes?)
IA158;Real Time Systems
PV021;Neuronové sítě

Logistic function

Cluster size: 4
IB031;Úvod do strojového učení
IA168;Algorithmic game theory
IA158;Real Time Systems
PV021;Neuronové sítě

Cluster size: 3
IB101;Úvod do logiky
IB102;Automaty, gramatiky a složitost
IB000;Matematické základy informatiky

Cluster size: 5
IA080;Seminar on Knowledge Discovery
PV115;Laboratoř dobývání znalostí
PV056;Strojové učení a dobývání znalostí
IA008;Computational Logic
PA164;Strojové učení a přirozený jazyk

Future?

- Some improvement suggestions:
  - Downloading grades and implement proper metric
  - Jaccard Index can be matched with some more advanced and better statistic methods
  - More cluster algorithms?

Thank you for your attention

Discussion helper I

- 6 letters, PV080-PV017 and PV080-IA101
  - TF-IDF 0,319536851569 0,013146465743 3
  - JI syllabus 0,148648648649 0,011235955056 2
  - JI objectives 0,149253731343 0,017391304347 8
- Basic stemming, same subjects
  - TF-IDF 0,227518352002 0,003108455845 97
  - JI syllabus 0,121495327103 0,046296296296 3
  - JI objectives 0,148936170213 0,030303030303

Discussion Helper II

- Threshold 3.0 3.5 4.0 4.5 5.0
  - Biggest cluster 11 23 54 127 350
  - Davies-Bouldin 0.01 0.031 0.078 0.201 0.416
  - Dunn 0.015 0.014 0.014 0.013 0.224
- Threshold 0.35 0.45 0.55 0.65 0.75 0.85 0.95
  - Biggest cluster 11 20 52 106 133 150 418
  - Davies-Bouldin 0.007 0.016 0.046 0.089 0.120 0.146 0.563
  - Dunn 0.107 0.107 0.102 0.1 0.191 0.409 0.546