annif tutorial

Data sets for this tutorial

yso-nlf and stw-zbw

1640
THE NATIONAL
LIBRARY
OF FINLAND

ZBW
Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

Data set of the National Library of Finland (NLF)

The folder yso-nlf in our GitHub repository contains

- the trilingual General Finnish Ontology YSO plus YSO-Places
- a training data set constructed from metadata records from the Finna.fi discovery service, and
- 2.066 English language Master's and doctoral theses published in the years 2010 to 2017 from the University of Jyväskylä (JYX repository).

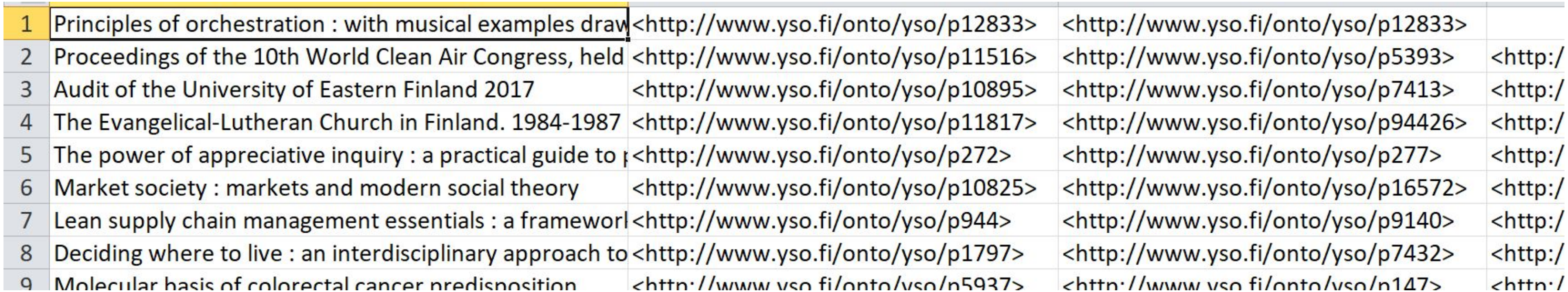
Data set of ZBW

The folder stw-zbw contains

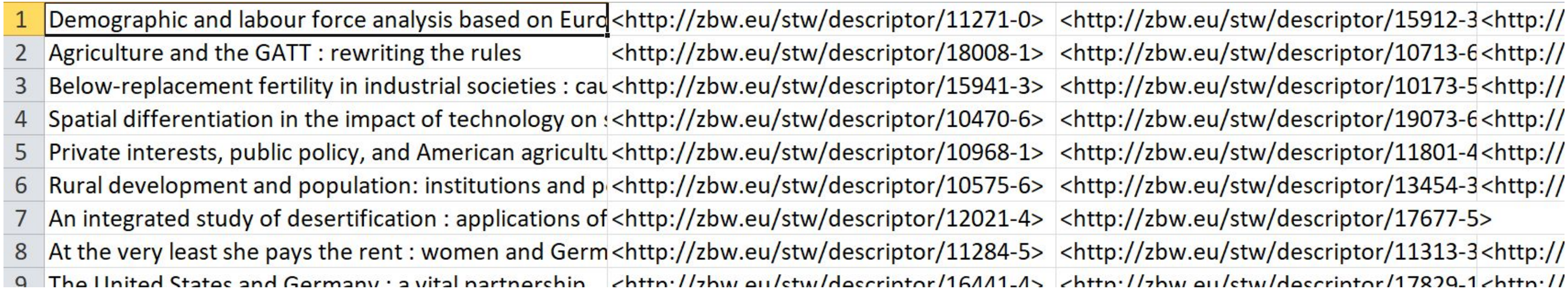
- the STW thesaurus for economics
- a training data set constructed from metadata records from the EconBiz discovery service
- 4.190 working papers in economics from the ZBW open access repository EconStor

Contents of short text training data sets

Title **Subjects (URIs)**

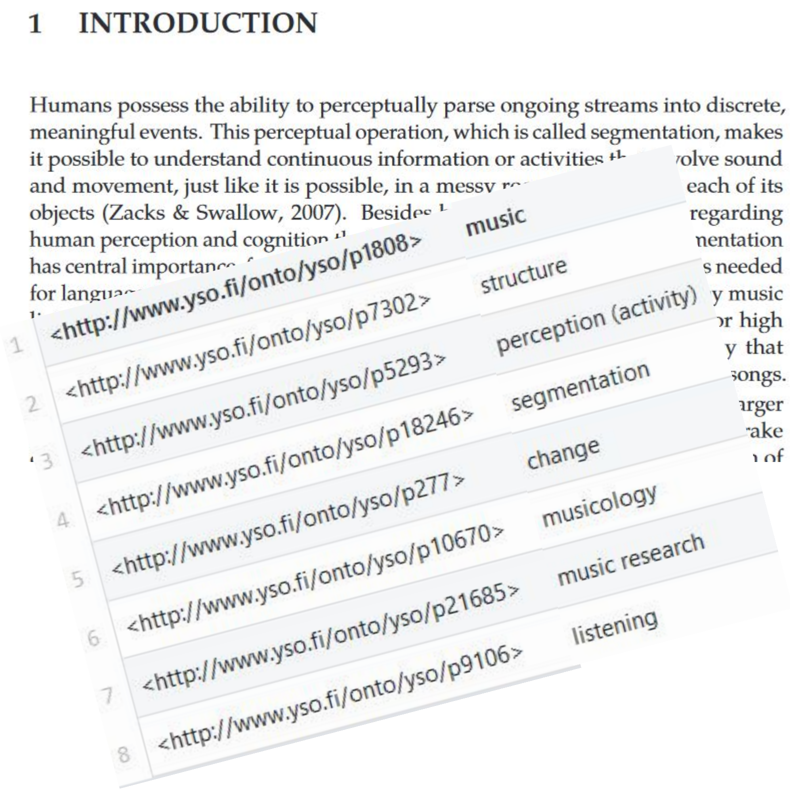


1	Principles of orchestration : with musical examples draw	<http://www.yso.fi/onto/yso/p12833>	<http://www.yso.fi/onto/yso/p12833>	
2	Proceedings of the 10th World Clean Air Congress, held	<http://www.yso.fi/onto/yso/p11516>	<http://www.yso.fi/onto/yso/p5393>	<http:/
3	Audit of the University of Eastern Finland 2017	<http://www.yso.fi/onto/yso/p10895>	<http://www.yso.fi/onto/yso/p7413>	<http:/
4	The Evangelical-Lutheran Church in Finland. 1984-1987	<http://www.yso.fi/onto/yso/p11817>	<http://www.yso.fi/onto/yso/p94426>	<http:/
5	The power of appreciative inquiry : a practical guide to	<http://www.yso.fi/onto/yso/p272>	<http://www.yso.fi/onto/yso/p277>	<http:/
6	Market society : markets and modern social theory	<http://www.yso.fi/onto/yso/p10825>	<http://www.yso.fi/onto/yso/p16572>	<http:/
7	Lean supply chain management essentials : a framewor	<http://www.yso.fi/onto/yso/p944>	<http://www.yso.fi/onto/yso/p9140>	<http:/
8	Deciding where to live : an interdisciplinary approach to	<http://www.yso.fi/onto/yso/p1797>	<http://www.yso.fi/onto/yso/p7432>	<http:/
9	Molecular basis of colorectal cancer predisposition	<http://www.yso.fi/onto/yso/p5937>	<http://www.yso.fi/onto/yso/p147>	<http:/



1	Demographic and labour force analysis based on Euro	<http://zbw.eu/stw/descriptor/11271-0>	<http://zbw.eu/stw/descriptor/15912-3	<http://
2	Agriculture and the GATT : rewriting the rules	<http://zbw.eu/stw/descriptor/18008-1>	<http://zbw.eu/stw/descriptor/10713-6	<http://
3	Below-replacement fertility in industrial societies : cau	<http://zbw.eu/stw/descriptor/15941-3>	<http://zbw.eu/stw/descriptor/10173-5	<http://
4	Spatial differentiation in the impact of technology on	<http://zbw.eu/stw/descriptor/10470-6>	<http://zbw.eu/stw/descriptor/19073-6	<http://
5	Private interests, public policy, and American agricultu	<http://zbw.eu/stw/descriptor/10968-1>	<http://zbw.eu/stw/descriptor/11801-4	<http://
6	Rural development and population: institutions and p	<http://zbw.eu/stw/descriptor/10575-6>	<http://zbw.eu/stw/descriptor/13454-3	<http://
7	An integrated study of desertification : applications of	<http://zbw.eu/stw/descriptor/12021-4>	<http://zbw.eu/stw/descriptor/17677-5>	
8	At the very least she pays the rent : women and Germ	<http://zbw.eu/stw/descriptor/11284-5>	<http://zbw.eu/stw/descriptor/11313-3	<http://
9	The United States and Germany : a vital partnership	<http://zbw.eu/stw/descriptor/16441-4>	<http://zbw.eu/stw/descriptor/17829-1	<http://

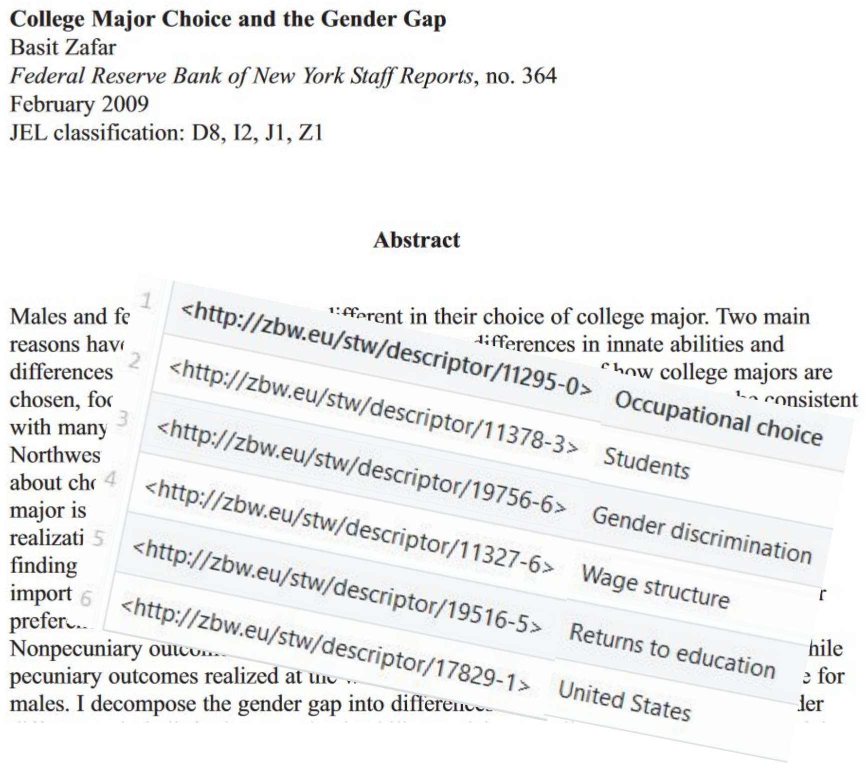
Contents of fulltext training data sets



1 INTRODUCTION

Humans possess the ability to perceptually parse ongoing streams into discrete, meaningful events. This perceptual operation, which is called segmentation, makes it possible to understand continuous information or activities th volve sound and movement, just like it is possible, in a messy r each of its objects (Zacks & Swallow, 2007). Besides h regarding human perception and cognition th mentation has central importance f s needed for langua y music r high y that songs. arger ake of

<http://www.yso.fi/onto/yso/p1808>	music
<http://www.yso.fi/onto/yso/p7302>	structure
<http://www.yso.fi/onto/yso/p5293>	perception (activity)
<http://www.yso.fi/onto/yso/p18246>	segmentation
<http://www.yso.fi/onto/yso/p277>	change
<http://www.yso.fi/onto/yso/p10670>	musicology
<http://www.yso.fi/onto/yso/p21685>	music research
<http://www.yso.fi/onto/yso/p9106>	listening



College Major Choice and the Gender Gap
Basit Zafar
Federal Reserve Bank of New York Staff Reports, no. 364
February 2009
JEL classification: D8, I2, J1, Z1

Abstract

Males and fe ifferent in their choice of college major. Two main reasons hav differences in innate abilities and differences how college majors are chosen, fo be consistent with many Northwes about ch major is realizati finding import r preferen Nonpecuniary outco hile pecuniary outcomes realized at the e for males. I decompose the gender gap into differenc der

<http://zbw.eu/stw/descriptor/11295-0>	Occupational choice
<http://zbw.eu/stw/descriptor/11378-3>	Students
<http://zbw.eu/stw/descriptor/19756-6>	Gender discrimination
<http://zbw.eu/stw/descriptor/11327-6>	Wage structure
<http://zbw.eu/stw/descriptor/19516-5>	Returns to education
<http://zbw.eu/stw/descriptor/17829-1>	United States

Both data sets at a glance

	vocabulary **(languages)** **(#concepts, terms)**	**short texts training docs**	**fulltexts** **(#; train, validate, test)**
NLF	YSO version 2019.3 Cicero (Finnish, Swedish, English) 32.265 concepts, 168.456 terms	~2 Mio. (~100.000 for testing)	Master’s & doctoral theses (2.066; 1.417, 349, 300)
ZBW	STW version 9.06 (German, English) 5.746 concepts, 32.272 terms	~1 Mio. (~100.000 for testing)	articles / working papers (4.190; 2.937, 628, 625)

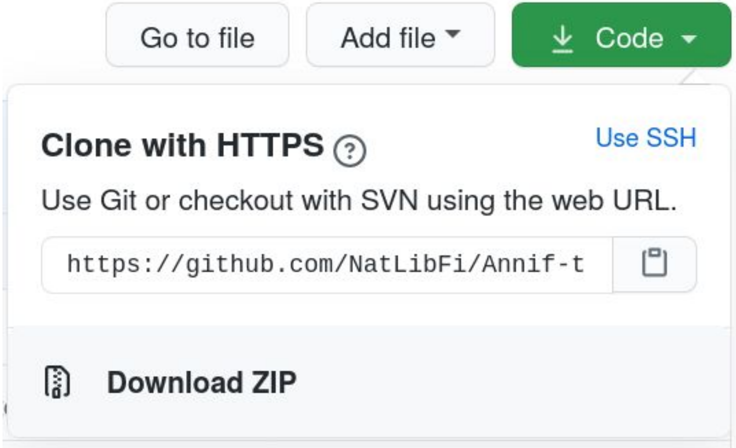
Choose one data set and use it for the rest of the tutorial

Getting the data sets

Step 1. Download the Annif-tutorial files

You need to download the data sets from the Annif-tutorial GitHub repository
- unless you are using the VirtualBox image, which already contains it

Download the zip file from https://github.com/NatLibFi/Annif-tutorial and extract it
- or if you know how to use git, you can just clone the repository!



Clone with HTTPS

Use Git or checkout with SVN using the web URL.

https://github.com/NatLibFi/Annif-t

Step 2. Download and convert the fulltext files

We cannot redistribute the fulltext documents for copyright reasons.

They must be downloaded separately from JYX/EconStor and converted to text files. This is automated using a Makefile, but it takes some time (30+ minutes) to run.

For the yso-nlf data set:

```
cd data-sets/yso-nlf/docs/
make -j4 -k
cd -
```

For the stw-zbw data set:

```
cd data-sets/stw-zbw/docs/
make -j4 -k
cd -
```

30 minutes later