Class 01: Getting started

September 13, 2017

Why are you taking this course?

Either:

- You don't know programming but are eager to learn, or
- It's a requirement for your degree

Good news!

- Programming is fun
- Programming improves all aspects of human experience
- Programming will make your life easier

More good news!

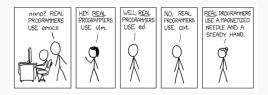
All the examples in this course are based on linguistic problems

Prerequisites

Stuff you need before you begin:

- A UNIX-compatible system (GNU/Linux, *BSD, Mac/OS)
- A text editor
- An installation of Python Python 3.0 or higher!

How to choose a text editor:



Honestly, use something other people (programmers) you know use.

Argh but what if I have Windows™

I have no idea about Windows

To be safe, install a Virtual Machine (e.g. VirtualBox) and a flavour of ${\sf GNU/Linux}, \ {\sf e.g.}$ Ubuntu.



On your own:

- A search engine such as GoogleTM, YandexTM or DuckDuckGoTM
- The fine Python documentation: http://docs.python.org
- Internet Relay Chat: http://webchat.freenode.net
- Stack Overflow: https://stackoverflow.com

Ack ma: In class 518 or in the corridor

ASK IIIe.	III Class, 510 of III the Corndor	(1114)
	#hseling on irc.freenode.net	(IRC)
	https://vk.com/id138461818	(VK)
	francis.tyers@gmail.com	(Hangouts)

(IDI)



On your own:

- A search engine such as GoogleTM, YandexTM or DuckDuckGoTM
- The fine Python documentation: http://docs.python.org
- Internet Relay Chat: http://webchat.freenode.net
- Stack Overflow: https://stackoverflow.com

Ask me:	In class, 518 or in the corridor	(IRL)
	<pre>#hseling on irc.freenode.net</pre>	(IRC)
	https://vk.com/id138461818	(VK)
	francis.tyers@gmail.com	(Hangouts)

```
1:32] == - 《朱儒> no.
1:33] == - 〈朱儒> no.
1:33] == - 〈朱儒> old saying...
1:33] == - 〈Fsociety> then let's destroy it now
1:33] == - 〈Fsociety> can we run the scripts in 30 seconds?
1:34] == - 〈朱儒> you misunderstood
1:34] == - 〈Fsociety> misunderstood what?
1:34] == - 〈朱儒> — Mode #da70_9RnPjm [+* lce47ks89@gateway/web/fre
```

On your own:

- A search engine such as Google[™], Yandex[™] or DuckDuckGo[™]
- The fine Python documentation: http://docs.python.org
- Internet Relay Chat: http://webchat.freenode.net
- Stack Overflow: https://stackoverflow.com

Ask me:	In class, 518 or in the corridor	(IRL)
	#hseling on irc.freenode.net	(IRC)
	https://vk.com/id138461818	(VK)
	francis.tyers@gmail.com	(Hangouts)



On your own:

- A search engine such as Google[™], Yandex[™] or DuckDuckGo[™]
- The fine Python documentation: http://docs.python.org
- Internet Relay Chat: http://webchat.freenode.net
- Stack Overflow: https://stackoverflow.com

francis.tyers@gmail.com

Ask me:	In class, 518 or in the corridor	(IRL)
	#hseling on irc.freenode.net	(IRC)
	https://vk.com/id138461818	(VK)

Structure of the course

https://ftyers.github.io/079-preproc-ling-dann/index.html

Class	Topic	Class	Topic
1	Command line	7	Project work
2	Segmenter	8	Project work
3	Tokeniser	9	Project work
4	Transliterator	10	Project work
5	Language model	11	Project work
6	Tagger	12	Presentations

Pipeline

A typical basic NLP pipeline looks like the following:

```
sentence segmenter | tokeniser | tagger | parser
```

- segmenter: takes a paragraph and gives sentences
- tokeniser: takes a sentence and gives list of tokens
- tagger: gives every token a morphosyntactic tag
- parser: takes a tagged sentence and gives a parse tree

During the first six classes you will be implementing basic versions of the first three modules.

Projects

For the remaining six classes you will work on:

- A small software project
- Something that you are excited about

For inspiration, you could:

- Perform some quantitative linguistic experiment
- Implement a program to convert between formats
- Write a scraper for some online language data
- Implement a simple machine learning solution to a problem

You will need to decide by the 5th class, if you are unsure, talk to me

Marking scheme

Details on the course page.

Marking

- 40% Project
 10% Exam
- 20% Practicals
- 20% Practicals
 20% Homework
- 10% Active participation

10% Active participation

Project: The project will encompass all of the class work and homework for the last six classes. You should start thinking from the first class what you might be interested in working on. If you cannot come up with any ideas, then I will give a number of options, or come and talk to me. The project should be substantial and test and expand your knowledge in some way. It should contain an evaluation component, either for efficiency of implementation or in terms of accuracy for some task. It will include a short (maximum ten minute) presentation to be done on the last day of class. One of the most important sapects of programming is learning to use the computer to scratch an itch 'удовлетворить личное желание' the project will ensure you are able to do that.

Exam: There will be a short exam to test you on what you have learnt in the course. There will be a number of multiple choice questions and a programming assignment to complete on paper. As programming is a subject that is more suited to practical evaluation the exam has a concomitantly low contribution to your final mark.

Practicals: Most of the course will be made up of practical sessions. I will evaluate your progress after each session.

Homework: Homework will be submitted through Github, and will need to be completed before the following lesson. Your Github repository should be called 2017-osnov_programm and have the following subdirectories: corpus for your (sub-)corpus from Wikipedia, and project for your project work.

Active practicipation: Beyond simply showing up, I encourage you to contribute to discussions by asking questions, answering quicipations, making relevant comments, belipting classmates and asking for help with in-class questions, are not supported by the process of the pro

tl;dr Most of the final mark is from the class work and project.

What we are going to do today

First things first:

- Make sure you have Python installed
- Set up Github accounts
- Install a text editor
- Work with the shell

Then second things:

- Choose a language
 - For purposes of speed, choose one with <= 500,000 articles
- Download the Wikipedia in that language
- Extract the text from Wikipedia

Check your Python installation

Open a terminal and type python3 and press return ${\ensuremath{\text{@}}}$.

```
$ python3
Python 3.5.2+ (default, Aug 5 2016, 08:07:14)
[GCC 6.1.1 20160724] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

If you don't have Python installed, install it now.

Github

All practical work will be stored and submitted through GitHub.

If you don't already have an account:

- Go to https://github.com/join
- Fill in the information
- Click "Create an account"
- Choose "Unlimited public repositories for free."
- Skip the next part.

Setup the directory structure

In your browser:

- First make a repository, call it 2017-preproc-ling-dann
- Choose 'Initialise this repository with a README'
- Click 'Clone or download' and copy the link

In the terminal:

\$ git clone https://github.com/XXXXXX/2017-preproc-ling-dann.git

\$ cd 2017-preproc-ling-dann

\$ mkdir corpus project

Where XXXXXX is your GitHub username.

Text editor

There are about 100500 text editors ...

I tried to get a definitive answer on which is the best text editor by asking your fellow students I know which one they use ...

- Sublime: +++
- TextWrangler: +
- Vim: +
- Atom: +
- Notepadpp: +
- Emacs:

Unfortunately there were nearly as many favourites as students ...

Wikipedia as a corpus/1



Wikipedia makes a great¹ corpus:

- Free to use and distribute
- Very many languages 295 at the last count

¹Well, great in some respects

				1.00	0 000+				
Deutsch	Español	Italiano	日本語	Polski	Sinugboanong	Svenska	Winaray		
English	Français	Nederlands		Русский	Binisaya	Tiếng Việt			
				100	000+				
العربية	Беларуская	Ελληνικά	한국어	עברית	Bahaso	Oʻzbekcha /	Simple English	Suomi	اردر
Azerbaycanca	(Акадэмічная)	Esperanto	Հայերեն	dafaanggaa	Minangkabau	Узбекча	Slovenčina	astrio	Volapük
Български	Català	Euskara	8-8	Latina	Norsk (Bokmål ·	Português	Slovenščina	ການາໃນນ	中文
Bân-lâm-gú /	Cestina	فارسي	Hrvatski	Lietuviu	Nynorsk)	Казакша /	Српски / Srpski	Türkçe	
Hô-ló-oê	Dansk	Galego	Bahasa Indonesia	Magyar	Нохчийн	قاز اقشا / Qazaqşa	Srpskohrvatski /	Українська	
	Eesti	9-		Bahasa Melayu		Română	Српскохрватски		
				10	000+				
Afrikaans	Беларуская	Frysk	Ирон аевзаг		BD100000	नेपाल भाषा	Plattdüütsch	Sec	so small Bas
Afrikaans Alemannisch	(Тарашкевіца)	Gaeilge	ирон æвзаг Islenska	Кырык Мары Latviešu	मत्तरी मत्तरी	नेपाली	Runa Simi	Basa Sunda	Ugi Das
Aremannisch Artici	(Tapaшkesiца) तिवृत्तिका प्रतिनृति	Gäidhlig	Jawa	Lätzebuergesch	მარგალური	Nnapulitano	Cymraeg	Kiswahili	Vèneto
Aragonés	Boarisch	ગુજરાતી	BEIER	Limburgs		Occitan	चंत्कृतम्	Tagalog	Walon
	Bosanski	Hornioserbsce	Kreyöl Ayisyen	Lumbaart	مصری	Occitan			WHIOII
Asturianu con	Brezhoneg	Ido	Kreyor Ayisyen کوردی / Kurdî	मेखिली	مازرونی Mina-děna-naû	पंतरवी (वारामधी)	Caxa Тыла Scots	Татарча / Tatarça Вкиги	Yorùbá
	Чавашла	Ilokano		Македонски	мінд-аенд-пду Монгол		Shqip	Точикй	御紙
Basa Banyumasan Башж ортса	чавашла Føroyskt	Interlingua	کوردیی ناوهندی Кыргызча	македонски Malagasy	МОНГОЛ	پنجابی (شاء مکھی) Piemontèis	Sicilianu	Точики نة, كحه	™as Zemaitėška
вашкортса	roroyski	interingua	KBDI BI348			Plemonteis	Sicilianu	نوردجه	Zemaneska
				1	+000				
Bahsa Acèh	मोजपुरी	Diné Bizaad	明語	Kinyarwanda	لزرى شومالي	Nouormand /	Picard	Sesotho sa Leboa	Тыва дыл
Адыгэбзэ	Bikol Central	Dolnoserbski	Hak-kā-fa / 客家話	Коми	Luganda	Normaund	Къарачай-	ChiShona	Удмурт
Ænglisc	Bislama	Emigliàn-	Хальмг	Kongo	Malti	Novial	Малкъар	سنڌي	دؤيدؤرجه
An ₃ cya	tel april	Rumagnöl	ھۇشا / Hausa	कोकमी / Konknni	文言	Олык Марий	Qaraqalpaqsha	Ślūnski	Vepsän
Armäneashce	Буряад	Эрзянь	'Ōlelo Hawai'i	ພາສາລາວ	Reo Má'ohi	पत्रशिष	Qırımtatarca	Soomaaliga	Vöro
Arpitan	Chavacano de	Estremeñu	Igbo	Dzhudezmo /	Māori	पाळि	Ripoarisch	Sranantongo	West-Vlams
المعدسا	Zamboanga	Fiji Hindi	Interlingue	לאדינו	Mirandés	Pangasinán	Rumantsch	Taqbaylit	Wolof
Avane'e	Corsu	Furlan	Kalaallisut	Лакку	Мокшень	Papiamentu	Русиньскый	Tarandine	吳語
Авар	Cuengh	Gaelg	Kapampangan	Лезги	Nähuatlahtölli	يښتو	Язык	Tetun	Zazaki
Aymar	Deitsch	Gagauz	Kaszébsczi	Liguru	Dorerin Naoero	Перем Коми	Sámegiella	Tok Pisin	Zeėuws
Bahasa Banjar	وقرشت	Gīkūyū	Kernewek	Lingála	Nedersaksisch	Pfälzisch	Sardu	faka Tonga	
		گيلکي		lojban	Nordfriisk		Seeltersk	Türkmençe	
				1	00+				
Akan	Evegbe	Iñupiak	Молдовеняска	Norfuk / Pitkern	gen.	Gagana Sâmoa	Словѣньскъ /	GWA	chiTumbuka
Bamanankan	Fulfulde	كفنيزي	Na Vosa Vaka-Viti	Afaan Oromoo	Romani	Sängö	SASPAPASSIA	Tséhesenéstsestots	e Twi
Chamoru	POTISH	Latgaļu	Něhiyawéwin /	Ποντιακά	Kirundi	Sesotho	SiSwati	Tshivenga	isiXhosa
Chichewa	Δα ⁰ Ω⊃ ^c / Inuktitul		₩ΔΥ∇Δ°			Setswana	9-mc/2	Xitsonga	isiZulu

Not on Wikipedia: Ainu, Chukchi, Dargwa, Khanty, Udi

```
Adyghe · Avar · Bambara · Bashkir · (Berber) · Breton · Chuvash · (East Caucasian) · Finnish · Hungarian · Kabyle · (Khoisan) · Komi · Lezgian · (Mande) · Mari · Mordvin · Rusyn · (Slavic) · Tatar · Udmurt · Yiddish
```

Too big: ${}^{?}$ English $\cdot {}^{?}$ French $\cdot {}^{?}$ German $\cdot {}^{?}$ Italian $\cdot {}^{?}$ Japanese $\cdot {}^{?}$ Polish $\cdot {}^{?}$ Russian $\cdot {}^{?}$ Spanish

Wikipedia as a corpus/3

Deliberately vague steps:

- Use your search engine to find where Wikipedia keeps its 'dumps'.
- Find the language code of the language you are interested in
- Download the dump for the language you are interested in
 - Tip 1: You're looking for a 'Database backup dump'
 - Tip 2: The filename will include pages-articles.xml.bz2
- Find WikiExtractor on the Apertium Wiki
- Run WikiExtractor on the dump file you downloaded.