Class 01: Getting started

September 11, 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 will make your life easier

More good news!

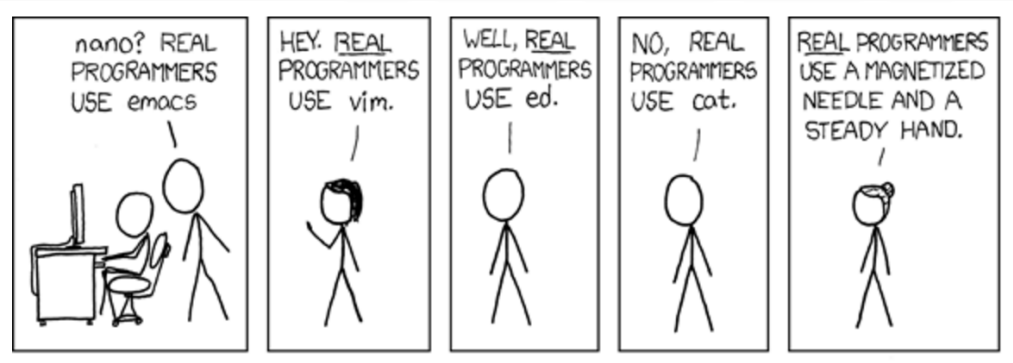
- All the examples in this course are based on linguistic problems

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™

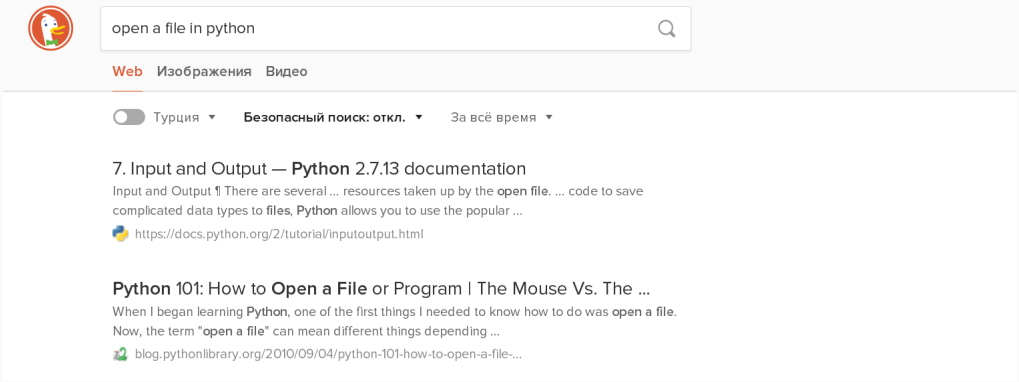
I have no idea about Windows

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

Installation instructions:

`http://wiki.apertium.org/wiki/Apertium_VirtualBox`

How to get help

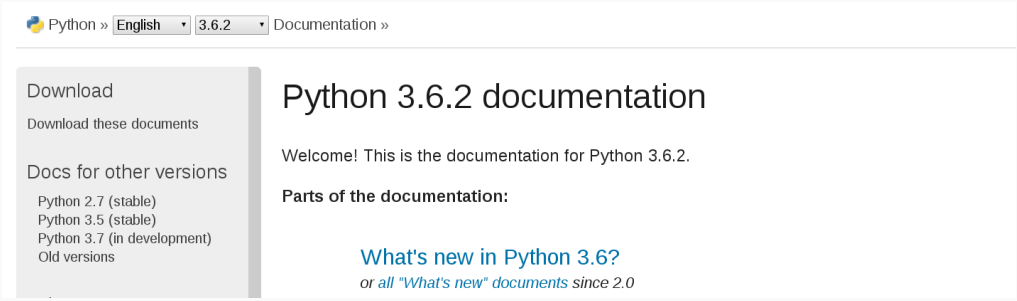


On your own:

- 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)

How to get help

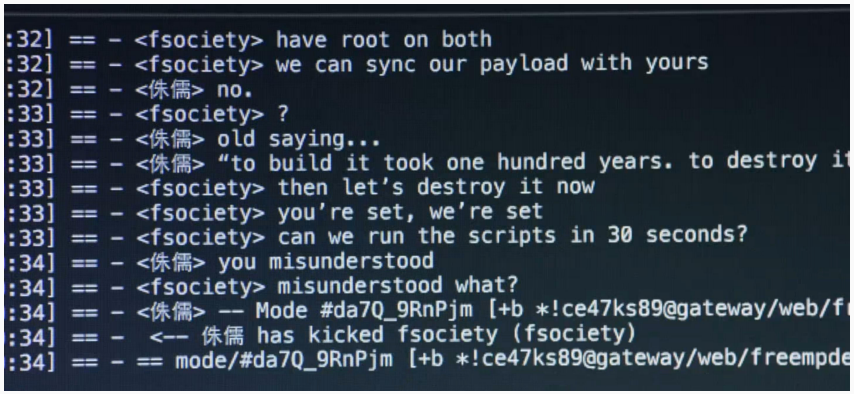


On your own:

- 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)

How to get help



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)

How to get help

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)

Structure of the course

`https://ftyers.github.io/079-osnov-programm/index.html`

Class	Topic	Class	Topic
1	Command line	5	Tagger
2	Segmenter	6	*Project work*
3	Tokeniser	7	*Project work*
4.1	Transliterator	8	*Project work*
4.2	Language model	–	–

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
- 25% Practicals
- 25% Homework
- 10% Active participation

Project: The project will encompass a good proportion of the class time and homework for the last three 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 non-trivial 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. One of the most important aspects 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.

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

Homework: Homework that isn't just reading 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. If you finish all practical work in a session, you can start on the homework.

Active participation: Beyond simply showing up, I encourage you to contribute to discussions by asking questions, answering questions, making relevant comments, helping classmates and asking for help with in-class activities, etc. There are no stupid questions — I want to make sure everyone grasps the concepts, and many are not as straightforward as they may first seem (or as I think they are). You are also expected to have read any assigned readings before class.

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 ⏎ .

```
$ 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-osnov-programm`
- 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-osnov-programm.git

$ cd 2017-osnov-programm

$ mkdir corpus project
```

Where XXXXXX is your GitHub username.

Text editor

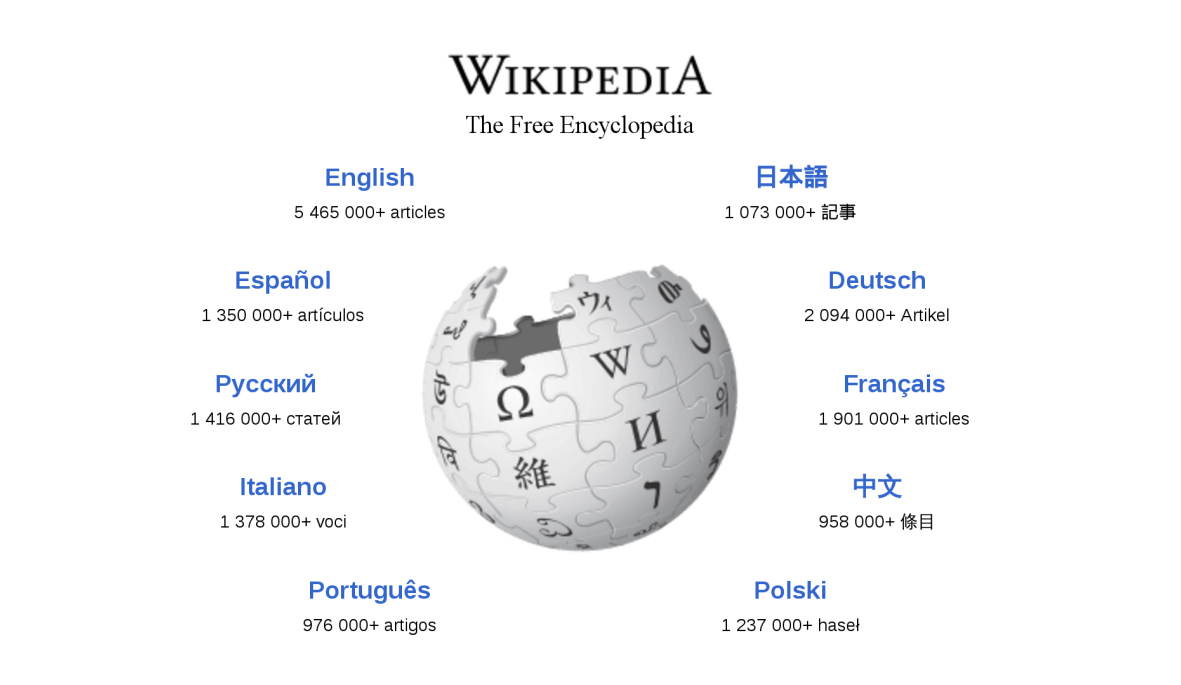
There are about 100500 text editors ...
I tried to get a definitive answer on which is the best text editor by asking your fellow students I know which one they use ...

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

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

† Not free/open-source software. :'(

Wikipedia as a corpus/1



Wikipedia makes a great[1] corpus:

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

[1] Well, great in some respects

文A Read Wikipedia in your language ^

1 000 000+

Deutsch · English · Español · Français · Italiano · Nederlands · 日本語 · Polski · Русский · Sinugboanong Binisaya · Svenska · Tiếng Việt · Winaray

100 000+

العربية · Azərbaycanca · Български · Bân-lâm-gú / Hô-ló-oē · Беларуская (Акадэмічная) · Català · Čeština · Dansk · Eesti · Ελληνικά · Esperanto · Euskara · فارسی · Galego · 한국어 · Հայերեն · हिन्दी · Hrvatski · Bahasa Indonesia · עברית · ქართული · Latina · Lietuvių · Magyar · Bahasa Melayu · Bahaso Minangkabau · Norsk (Bokmål · Nynorsk) · Нохчийн · Oʻzbekcha / Ўзбекча · Português · Қазақша / Qazaqşa / قازاقشا · Română · Simple English · Slovenčina · Slovenščina · Српски / Srpski · Srpskohrvatski / Српскохрватски · Suomi · தமிழ் · ภาษาไทย · Türkçe · Українська · اردو · Volapük · 中文

10 000+

Afrikaans · Alemannisch · አማርኛ · Aragonés · Asturianu · বাংলা · Basa Banyumasan · Башҡортса · Беларуская (Тарашкевіца) · বিষ্ণুপ্রিয়া মণিপুরী · Boarisch · Bosanski · Brezhoneg · Чӑвашла · Føroyskt · Frysk · Gaeilge · Gàidhlig · ગુજરાતી · Hornjoserbsce · Ido · Ilokano · Interlingua · Ирон æвзаг · Íslenska · Jawa · [illegible] · Kreyòl Ayisyen · Kurdî / كوردی · کوردیی ناوەندی · Кыргызча · Кырык Мары · Latviešu · Lëtzebuergesch · Limburgs · Lumbaart · मैथिली · Македонски · Malagasy · മലയാളം · मराठी · მარგალური · مصرى · مازِرونی · Mìng-dĕ̤ng-ngṳ̄ · Монгол · [illegible] · नेपाल भाषा · नेपाली · Nnapulitano · Occitan · ଓଡ଼ିଆ · ਪੰਜਾਬੀ (ਗੁਰਮੁਖੀ) · پنجابی (شاہ مکھی) · Piemontèis · Plattdüütsch · Runa Simi · Cymraeg · संस्कृतम् · Саха Тыла · Scots · Shqip · Sicilianu · සිංහල · Basa Sunda · Kiswahili · Tagalog · Татарча / Tatarça · తెలుగు · Тоҷикӣ · تۆرکجه · ᨅᨔ ᨕᨘᨁᨗ / Basa Ugi · Vèneto · Walon · ייִדיש · Yorùbá · 粵語 · Žemaitėška

1 000+

Bahsa Acèh · Адыгэбзэ · Ænglisc · Аԥсуа · Armãneashce · Arpitan · ܐܬܘܪܝܐ · Avañeʼẽ · Авар · Aymar · Bahasa Banjar · भोजपुरी · Bikol Central · Bislama · བོད་ཡིག · Буряад · Chavacano de Zamboanga · Corsu · Cuengh · Deitsch · ދިވެހިބަސް · Diné Bizaad · Dolnoserbski · Emiliàn–Rumagnòl · Эрзянь · Estremeñu · Fiji Hindi · Furlan · Gaelg · Gagauz · Gĩkũyũ · گیلکی · 贛語 · Hak-kâ-fa / 客家語 · Хальмг · Hausa / هَوُسَ · ʻŌlelo Hawaiʻi · Igbo · Interlingue · Kalaallisut · Kapampangan · Kaszëbsczi · Kernewek · [illegible] · Kinyarwanda · Коми · Kongo · कोंकणी / Konknni · ພາສາລາວ · Dzhudezmo / לאדינו · Лакку · Лезги · Líguru · Lingála · lojban · لۊری شومالی · Luganda · Malti · 文言 · Reo Māʼohi · Māori · Mirandés · Мокшень · Nāhuatlahtōlli · Dorerin Naoero · Nedersaksisch · Nordfriisk · Nouormand / Normaund · Novial · Олык Марий · অসমীয়া · पालि · Pangasinán · Papiamentu · پښتو · Перем Коми · Pfälzisch · Picard · Къарачай–Малкъар · Qaraqalpaqsha · Qırımtatarca · Ripoarisch · Rumantsch · Русиньскый Язык · Sámegiella · Sardu · Seeltersk · Sesotho sa Leboa · ChiShona · سنڌي · Ślůnski · Soomaaliga · Sranantongo · Taqbaylit · Tarandíne · Tetun · Tok Pisin · faka Tonga · Türkmençe · Тыва дыл · Удмурт · ئۇيغۇرچە · Vepsän · Võro · West-Vlams · Wolof · 吳語 · Zazaki · Zeêuws

100+

Akan · Bamanankan · Chamoru · Chichewa · Eʋegbe · Fulfulde · 𐌲𐌿𐍄𐌹𐍃𐌺 · ᐃᓄᒃᑎᑐᑦ / Inuktitut · Iñupiak · كٲشُر · Latgaļu · Молдовеняскэ · Na Vosa Vaka-Viti · Nēhiyawēwin / ᓀᐦᐃᔭᐍᐏᐣ · Norfuk / Pitkern · Afaan Oromoo · Ποντιακά · རྫོང་ཁ · Romani · Kirundi · Gagana Sāmoa · Sängö · Sesotho · Setswana · Словѣ́ньскъ / ⰔⰎⰑⰂⰡⰐⰠⰔⰍⰟ · SiSwati · ትግርኛ · ᏣᎳᎩ · Tsėhesenėstsestotse · Tshivenḓa · Xitsonga · chiTumbuka · Twi · isiXhosa · isiZulu

Other languages · Weitere Sprachen · Autres langues · Kompletna lista języków · 他の言語 · Otros idiomas · 其他語言 · Другие языки · Aliaj lingvoj · 다른 언어 · Ngôn ngữ khác

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 · ?French · ?German · ?Italian · ?Japanese · ?Polish · ?Russian · ?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.