

# CS918: LECTURE 1a

Introduction to Natural Language Processing

---

Arkaitz Zubiaga, 3<sup>rd</sup> October, 2018

## ABOUT THE MODULE: CS918

- **Lectures:** Mon (9am, LIB2) & Wed (10am, S0.21).
- **Seminars:** Mon (1pm, LIB2) (week 2 onwards).
- **Labs (weeks 3, 5, 7, 9, 10):**
  - Group 1: Mon 4pm (CS0.01).
  - Group 2: Thu 4pm (CS0.06).

## ABOUT THE MODULE: CS918

- **Assessment:**
  - 70% exam in May/June.
  - 30% of 2 assignments:
    - Assign. 1 (10%): deadline week 6.
    - Assign. 2 (20%): deadline week 11.

## AIMS OF THE MODULE: CS918

- Give fundamental **understanding of NLP methods** for processing linguistic data in textual form.
- Familiarisation with different **applications of NLP** (e.g. sentiment analysis).
- Give **skills to apply** state of the art **NLP methods on different types of text** (newswire, web, social media, scientific articles).

## BOOKS FOR THE MODULE

- Jurafsky, Daniel, and James H. Martin. 2009. **Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics**. 3rd edition.
- Bird Steven, Ewan Klein, and Edward Loper. **Natural Language Processing with Python**. O'Reilly Media, Inc., 2009.
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, **Introduction to Information Retrieval**. Cambridge University Press. 2008.

## PROGRAMMING IN THE MODULE

- We will be using **Python!**
  - Many packages available for text processing.



Natural Language Analysis  
with Python NLTK



## LECTURE 1: CONTENTS

- What is Natural Language Processing (NLP)?
- What are NLP areas and applications?
- Why is NLP challenging?
- Basic text processing with Regular Expressions.

## WHAT IS NATURAL LANGUAGE PROCESSING?

- NLP studies **computational methods** for processing and understanding **human language** data (e.g. English or Chinese).
- In this module, **we will focus on textual** rather than spoken language.



## WHY IS NLP IMPORTANT?

- **A lot of today's knowledge is written in texts**, even more so on the Internet, social media, emails.
  - We need automated means to process all that content!
- **Communication with chatbots and across languages** needs understanding of language.
- ...and many more.

## WHY IS NLP IMPORTANT?

- Is being increasingly used by companies, e.g.:



## WHY IS NLP IMPORTANT?

- And required for more and more jobs.



Natural Language Processing (NLP) ⓘ				
UK				
Location	UK	6 months to 27 Sep 2018	Same period 2017	Same period 2016
Rank		555	721	739
Rank change year-on-year		▲ +166	▲ +18	▲ +350
Permanent jobs citing Natural Language Processing		760	525	569

## BRIEF HISTORY OF NLP

- **1940s:** used mainly for machine translation.
- **1980s:** Gained momentum with a focus on computational grammars for the representation of meaning. Small corpora, mostly rule-based.
- **1990s:** Rapid expansion, large collections, Internet.
- **2000s:** Shift from computational grammars to statistical (machine learning).
- **2013-:** Largely influenced by Deep learning.

# NLP APPLICATIONS: QUESTION ANSWERING

Google

what is the capital of the united kingdom

All Images Maps News Videos More Settings Tools

About 8,720,000 results (1.01 seconds)

United Kingdom / Capital



London

# NLP APPLICATIONS: QUESTION ANSWERING



Google

what is the population of the capital of the united kingdom?

All Maps News Images Shopping More Settings Tools

About 12,100,000 results (0.82 seconds)

London / Population

**8.788 million**

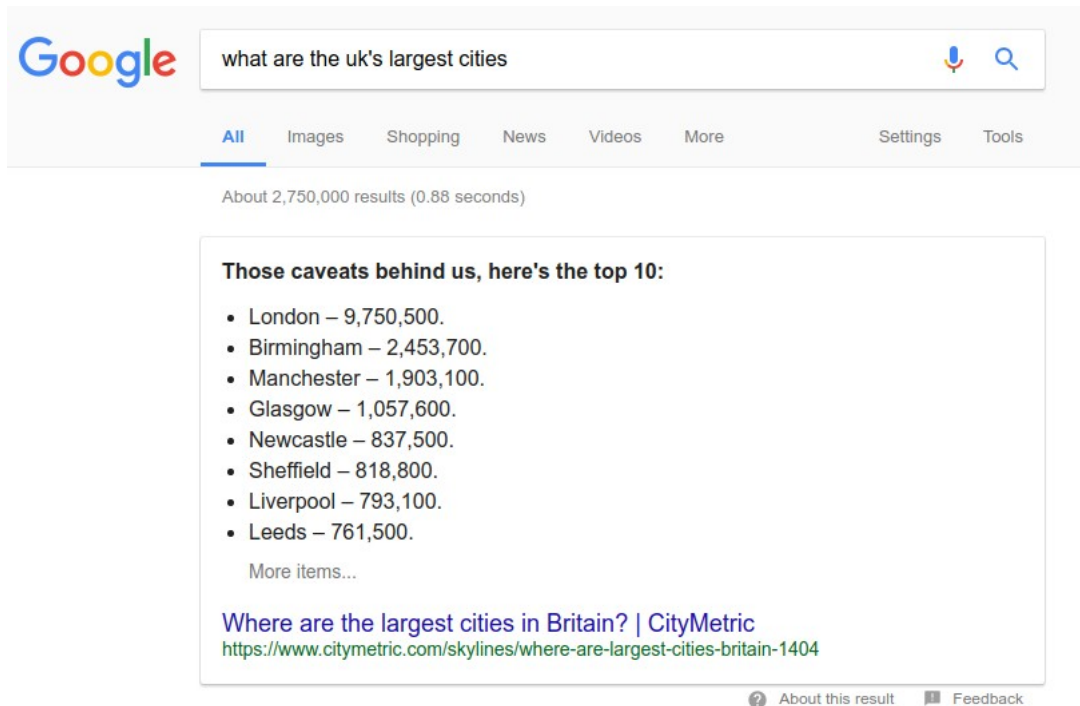
2016

People also search for

	New York City 8.538 million		United Kingdom 65.64 million		Paris 2.244 million
---	--------------------------------	---	---------------------------------	---	------------------------

Feedback

# NLP APPLICATIONS: QUESTION ANSWERING



Google

what are the uk's largest cities

All Images Shopping News Videos More Settings Tools

About 2,750,000 results (0.88 seconds)

**Those caveats behind us, here's the top 10:**

- London – 9,750,500.
- Birmingham – 2,453,700.
- Manchester – 1,903,100.
- Glasgow – 1,057,600.
- Newcastle – 837,500.
- Sheffield – 818,800.
- Liverpool – 793,100.
- Leeds – 761,500.

More items...

[Where are the largest cities in Britain? | CityMetric](https://www.citymetric.com/skylines/where-are-largest-cities-britain-1404)  
<https://www.citymetric.com/skylines/where-are-largest-cities-britain-1404>

About this result Feedback

## NLP APPLICATIONS: INFORMATION EXTRACTION

WARWICK

Subject: **meeting**

Date: 8<sup>th</sup> January, 2018

To: Arkaitz Zubiaga

**Event:** Meeting w/ Mike  
**Date:** 15 Jan, 2018  
**Start:** 10:00am  
**End:** 11:00am  
**Where:** A. Lovelace

Hi Arkaitz, we have finally scheduled the meeting.

It will be in the Ada Lovelace room, next Monday 10am-11am.

-Mike

Create new Calendar entry



## NLP APPLICATIONS: SENTIMENT ANALYSIS

**Booking.com**

"Very comfy bed, we stayed in a suite and it was lovely, very big and clean - had microwave, toaster and fridge. Good location right near the center, and has restaurants and shops across the road."

A Alix  
🇬🇧 United Kingdom

"we had a double room, but was to cold when we complaint about the heating not been working they move us to a suite, it was very nice and comfortable."

J Jose  
🇬🇧 United Kingdom

"The waitress at breakfast was very cheerful and friendly."

S Sandra  
🇬🇧 United Kingdom

"The bed was really comfy and the room nice and quiet"

S Susan  
🇬🇧 United Kingdom

# NLP APPLICATIONS: SENTIMENT ANALYSIS



"Very comfy bed, we stayed in a suite and it was lovely, very big and clean - had microwave, toaster and fridge. Good location right near the center and has restaurants and shops across the road."

A Alix  
United Kingdom

"we had a double room, but was to cold when we complaint about the heating not been working they move us to a suite, it was very nice and comfortable."

J Jose  
United Kingdom

"The waitress at breakfast was very cheerful and friendly."

S Sandra  
United Kingdom

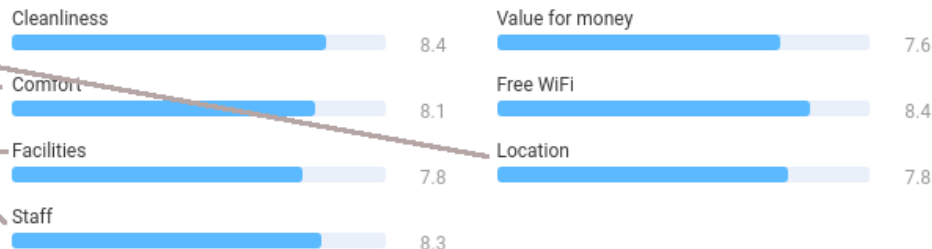
"The bed was really comfy and the room nice and quiet"

S Susan  
United Kingdom

## Guest reviews (1,625)

Real guests. Real stays. Real opinions. [Read more](#)

**8.0** Very good · 1,625 reviews ▾



# NLP APPLICATIONS: MACHINE TRANSLATION

WARWICK

光明网 gmw.cn

2016年1月2日 星期二

文化人, 天下事

English | 光明员工 | 邮箱

新闻 时政 国际 地方 时评 理论 党建 文化 科技 教育 经济 生活 旅游 养生 阅读 公益 食品 军事 法治 读图 专题 学术 历史 国学 卫生 科普 电视 女人 文娱 体育 乐跑 书画 文荟 药品

文艺评论 图片库 棋牌 留学 学术出版 中医 律师

核心价值百场讲坛 航拍中国征集活动

中宣机关学十九大精神 身边正能量摄影大赛

首都文化企业30强推选 图片库 光明图刊

荣宝斋青年书法篆刻展 网络举报监督专区

军事科技前沿

## 三论红船初心: “红船精神”的时代呼唤

[习近平: 弘扬“红船精神” 走在时代前列] [从“红船精神”看中国共产党的人民性特质] [再论红船初心]

光明图片

开启新时代中国特色大国外交壮阔征程

欧美同学会学习《习近平给莫斯科大学中国留学生的回信》

2017年反腐倡廉十件大事 保持定力, 将反腐进行到底

韩向前提议举行高层会谈 中方愿双方互释积极信息

央视独家记录仪仗队升旗前3小时 升旗仪式有七大变化

中央气象台发布暴雪黄色预警: 甘肃湖北等地有大雪

管清友: 中央经济工作会议的十大亮点

光明日报副总编辑陆高: 与时俱进的“娘家人”

天文学家发布包含3.5亿颗恒星和星系信息的星表

每年被吃掉1500亿片 阿司匹林真是万金油?

光明网评论员

中国改革开放40年: 前行是你的宿命

时评

有必要让“公加”

为中国人民进发出的创造伟力喝彩

穆轩宇: 爱国主义始终是主旋律

太平间不应成为殡葬公司的跑马场

流感辟谣并不意味着不防流感

经典价值历久弥新

热点

红船初心

弘扬“红船精神” 走在时代前列

光明网 gmw.cn

Tuesday, January 2, 2018

Culture, World Affairs

News Local political international military rule of law map feature

Commentary theory of party building academic history of Chinese culture

Culture science and technology education health science TV

Bright employees

Economic life tourist woman entertainment sports

Health reading public food Reeborg painting Wenhui drugs

Literary Criticism Photo Gallery Chess Study Abroad Academic Publishing Chinese Medicine Lawyer

organs of the Top 30 Capital Culture Rong Baozhai young Horror video report Internet reporting

## 三论红船初心: “红船精神”的时代呼唤

[Xi Jinping: to promote the "Red Boat Spirit" at the forefront of the times] [From the "Red Boat Spirit" People look at the characteristics of the Chinese Communist Party: 1. As the heart and soul of the party]

光明图片

Open a Great Extrude of Great Powers with Chinese Characteristics in the New Era

European and American students will learn "Xi Jinping's reply to Chinese students at Moscow University"

In 2017, the 10 major tasks of combating corruption and

光明网评论员

40 years of reform and opening up

时评

It is necessary to get "openly and forbidden" out of dispute

Applause the Chinese people for creating great power

# NLP APPLICATIONS

mostly solved

## Spam detection

Let's go to Agra!



Buy V1AGRA ...



## Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

## Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

making good progress

## Sentiment analysis

Best roast chicken in San Francisco!



The waiter ignored us for 20 minutes.

## Coreference resolution

Carter told Mubarak he shouldn't run again.

## Word sense disambiguation (WSD)

I need new batteries for my *mouse*.



## Parsing

I can see Alcatraz from the window!

## Machine translation (MT)

第 13 届上海国际电影节开幕...



The 13<sup>th</sup> Shanghai International Film Festival...

## Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30



Party  
May 27  
add

WARWICK

still really hard

## Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

## Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

## Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose



Economy is good

## Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?



## WHY IS NLP CHALLENGING?

- Language is ambiguous, e.g. "Flying planes can be dangerous"
- What is actually meant?
  - It can be dangerous for a person to fly planes.
  - Planes that are flying in the air can be dangerous.

## WHY ELSE IS NLP CHALLENGING?

### non-standard English

We had a double room, but was  
**to** cold when we **complaint**

### segmentation issues

the London Euston-Birmingham  
New Street train

**is Euston-Birmingham a word?**

### idioms

a pain in the neck

throw in the towel

### neologisms

unfriend  
Retweet  
selfie

### world knowledge

Mary and Sue are sisters.  
Mary and Sue are mothers.

### tricky entity names

*Let It Be* is a good song...

They were listening to *One Direction*...