# Wikipedia

## Lexical Resources

December $4^{th}$, 2019

# Contents of the lecture

# Wikipedia & how to handle it

# Wikipedia

## In broad terms

- Wikipedia is a collaborative encyclopedia
- The English version of Wikipedia contains $\approx$ 5,977,089 content pages.
- Wikipedia exists in 297 languages (+ 10 inactive projects), cf. https://en.wikipedia.org/wiki/List_of_Wikipedias

**Wikipedia** (/ˌwɪkɪˈpiːdiə/ (listen) *wik-ih-PEE-dee-ə* or /ˌwɪkiˈpiːdiə/ (listen) *wik-ee-PEE-dee-ə*) is a multilingual online encyclopedia created and maintained as an open collaboration project[3] by a community of volunteer editors using a wiki-based editing system.[4] It is the largest and most popular general reference work on the World Wide Web,[5][6][7] and is one of the most popular websites ranked by Alexa as of October 2019.[8] It features exclusively free content and no commercial ads, and is owned and supported by the Wikimedia Foundation, a non-profit organization funded primarily through donations.[9][10][11][12]

# Wikipedia

## Content pages

- Wikipedia is an online encyclopedia, ie. "a reference work or compendium providing summaries of knowledge either from all branches or from a particular field or discipline."(def. from Wikipedia)
- In the case of Wikipedia, these summaries are known as Wikipedia 'pages' (more precisely 'content pages') or 'articles'.
- As is often the case with collaborative projects, the quality and reliability of pages **vary** (both across pages, across domains and across languages).

# Wikipedia

## Infoboxes and Lead sections

There are multiple ways of extracting information from a wikipedia page. The NLP community generally focuses on two specific elements

1. Lead Section: The few first sentences of an article, that broach over a subject and attempts to give both overview and introduction to the full contents of the page (cf. https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section).
2. Infobox: Wikipedia contains 'infoboxes', viz. tabulated data in attribute-value format presenting the key facts of the related page (cf. https://en.wikipedia.org/wiki/Infobox)

**French Republic**
*République française*

Flag

Emblem

**Motto:** "Liberté, Égalité, Fraternité" (French)
"Liberty, Equality, Fraternity"

**Anthem:** "*La Marseillaise*"
0:00 CC MENU

Location of France (dark green)
in the European Union (light green)

| | |
|---|---|
| **Capital** and largest city | Paris 48°51.4′N 2°21.05′E |
| **Official language** and national language | French[I] |
| **Government** | Unitary semi-presidential constitutional republic |
| • President | Emmanuel Macron |
| • Prime Minister | Édouard Philippe |
| **Legislature** | Parliament |
| • Upper house | Senate |
| • Lower house | National Assembly |
| **Establishment** | |
| • Current constitution | 4 October 1958 (61 years) |
| **Currency** | Euro (EUR) CFP franc (XPF) |
| **Date format** | dd/mm/yyyy (AD) |
| **Calling code** | +33[II] |
| **ISO 3166 code** | FR |
| **Internet TLD** | .fr[III] |

# Wikipedia

## Wikipedia for the developer

- All the content of all Wikipedias is available for download: https://dumps.wikimedia.org/
- There are many available guidelines regarding how Wikipedia articles are to be structured, in particular for wiki markup cf. https://en.wikipedia.org/wiki/Help:Wikitext
- There are many existing tools devoted to handling wiki markup: eg. the pip package `wikipedia`, the pip package `wptools` or the gensim `corpora.wikicorpus` subpackage.

# Elements of graph theory

# Elements of graph theory

## Wikipedia is a graph

We can consider that Wikipedia is a graph, where $V$ is the set of Wikipedia pages, and $E \subseteq V \times V$ is the set of internal links. If a page $p_i$ refers to another page $p_j$ using a internal link, then $\langle v_i, v_j \rangle \in E$: one page ($p_j$) can be accessed from another ($p_i$), hence one vertex $v_i$ is linked to another ($v_j$).

More formally:

- A graph is a pair $\langle V, E \rangle$ where $V$ is a set of vertices, or nodes, and $E \subseteq V \times V$ is a set of oriented edges linking vertices two by two.
- the number of edges $\langle v_x, v_i \rangle$ arriving at a given vertex $v_i$ is called the indegree of $v_i$
- the number of edges $\langle v_i, v_x \rangle$ starting from a given vertex $v_i$ is called the outdegree of $v_i$
- By definition, $\sum_{v_i \in V} \text{indegree } (v_i) = \sum_{v_i \in V} \text{outdegree } (v_i) = \#E$

# Elements of graph theory

## It's a small world after all

Many empirical graphs are 'small world' graphs or 'small world networks'. Intuitively, such graphs have few 'hubs' (vertices with very high degrees) and many 'neighbors' (vertices connected to a 'hub'). There's no absolute definition, but we generally employ the following:

- Let $N(v_i) = \{v_n \textit{ such that } \langle v_i, v_n \rangle \in E\}$, the 'outgoing neighborhood' of node $v_i$.
- Let $C(v_i) = \frac{\#\{\langle v_n, v_m \rangle \textit{ such that } v_n, v_m \in N(v_i) \ \wedge \ \langle v_n, v_m \rangle \in E\}}{k(k-1)}$ the 'local clustering coefficient' of node $v_i$, where $k = \#N(v_i)$.
- Let the length of the shortest path between $v_i$ and $v_j$

$$L(v_i, v_j) = \begin{cases} \infty & \textit{if } N(v_i) = \emptyset \\ 1 & \textit{if } \langle v_i, v_j \rangle \in E \\ 1 + \min\limits_{v_n \in N(v_i)} L(v_n, v_j) & \textit{otherwise} \end{cases}$$

- If $\bar{L} \propto \log \#V$ and $\bar{C}$ is sufficiently large ($>$ to what a random graph with the same number of vertices would have), the graph is 'small-word'.

One can think of 'small-world' structures as the analog for graphs of 'Zipf's Law' for word frequencies.

See also: https://en.wikipedia.org/wiki/Wikipedia:Getting_to_Philosophy

# Example applications

# What would you do with a wiki

## Some applications of Wikipedia

The main usage of encyclopediae at large is to ground NLP applications into the real world, by making use of the facts that they collect.

- Typically, an important branch of research has been dedicated in converting online encyclopediae (and Wikipedia in particular) into 'ontologies'.
- Other usages include deriving templates and models for converting raw data (attribute-values pairs or raw numerical data) into natural-language text.

# What would you do with a wiki

## Inducing ontologies

Ontologies are systematic definitions for classifying and structuring data. 'Ontology learning' is the task that focuses on inferring such systematic classifications from unstructured or partially structured data. A generally well established format for Ontologies is the Resource Description Framework proposed by the W3C.

- More specifically restricted to Wikipedia, the DBPedia initiative aims at extracting factual relations from Wikimedia pages and converting those to the RDF format.
- **Vast enterprise**: "1,445,000 persons, 735,000 places (including 478,000 populated places), 411,000 creative works (including 123,000 music albums, 87,000 films and 19,000 video games), 241,000 organizations (including 58,000 companies and 49,000 educational institutions), 251,000 species and 6,000 diseases."

# What would you do with a wiki

## Data-to-text NLG

Natural language generation ('NLG') anchored into factual data is one of the 'holy grails' of NLP.
Using Wikipedia:

- Retrieve the attribute-value pairs listed in the infobox
- Transform these pairs into vector representations, eg. using an embedding-based biaffine function $\mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ : $\vec{a} = e(a);\ \vec{v} = e(v);\ \vec{r} = B(\vec{a}, \vec{v})$ with $a$ the attribute, $v$ the value and $B$ the learned composition function. A simple concatenation followed by a linear layer could suffice. Another way of tackling this would be using copy pointer mechanisms.
- These representations can be thought of as a source, and the lead section as a target. Following that, any sequence-to-sequence algorithm can function. Attention-based models in particular will prove useful, as they should allow you to use the multiple source representations efficiently.