Dataset

原始資料來源是 http://archives.textfiles.com/ 其中資料主要用於自然語言處理培訓和測試，我們只需要處理其中的科學主題，它包含 150 多個屬於不同主題的文字檔案。為了方便資料處理，我們預先計算了這些檔案裡找到的關鍵字的 tf.ifd 值（可參考：https://ithelp.ithome.com.tw/articles/10214022 ），以便排名重要性。

Tasks

Task 1: Highlight words

第一個任務是視覺化個別檔案，透過視覺上強調單個關鍵字在文本和語料庫中的重要性；具體來說，產出是印出（print out）文本且改變單個關鍵字的顏色和字體粗細來呈現重要性。這邊的重要性定義是每個詞語相關的 tf.idf 值，程式上需要將檔案名稱作為 input，並以和原始文字檔相同的格式顯示文字，只需要在顏色和字體粗細上做改變。以下示意圖為 blackhol.txt 和 fusion.txt 的可能結果。

API: `python a4_highlights.py -d [--document] <name_of_document>`

Notes: 1) 如果你使用 Matplotlib，可參考 here 來修改文字的外觀；2) 如果使用 Bokeh，可使用 Div 達到同樣的效果。

B L A C K H O L E S I N S P A C E

There is much more to **black holes** than meets the eye. In fact, your eyes, even with the aid of the most advanced telescope, will never see a **black hole** in space. The reason is that the **matter** within a **black hole** is so dense and has so great a **gravitational** pull that it prevents even **light** from escaping.

Like other electromagnetic **radiation** (radio waves, infrared rays, ultraviolet radiation, X-rays, and **gamma** radiation), **light** is the fastest traveler in the Universe. It moves at nearly 300,000 **kilometers** (about 186,000 miles) per second. At such a speed, you could circle the **Earth** seven **times** between heartbeats.

If **light** can't **escape** a **black** hole, it follows that nothing else can. Consequently, there is no direct way to detect a **black** hole.

In fact, the principal evidence of the existence of **black holes** comes not from observation but from solutions to complex equations based on Einstein's **Theory** of General Relativity. Among other things, the calculations indicate that **black holes** may occur in a variety of sizes and be more abundant than most of us realize.

MINI BLACK HOLES

Some **black holes** are **theorized** to be nearly as old as the Big Bang, which is hypothesized to have started our **Universe** 10 to 20 billion years ago. The rapid early **expansion** of some parts of the dense hot **matter** in this nascent **Universe** is said to have so

FUSION PRINCIPLES

Under solar conditions (high temps of about 100 millions degrees C, 1 milion megabars pressure), H **atoms** fuse into He. Three **isotopes** of H exist:

H1 (P) **protium**
H2 (D) **deuterium**
H3 (T) tritium.

Protium reacts too slowly even in the sun so **deuterium** and **tritium** are used. Under solar conditions, the H **atoms** gain enough **kinetic** enery to overcome the **electrostatic repulsion** of their **positive** charges. The electrons which are normally found surrounding H nuclei have already been ionised. You have a plasma of **positive** nuclei. He is formed in a H-H reaction, releasing energy.

Sources of D and T

Heavy **water** (D2O) is present at 1 part in 6700 in normal tap water. You can **separate** the heavy water, and then obtain **deuterium** gas. D2 **gas** is obtained via electrolysis.

Tritium is radioactive, and is obtained via bombardment of Li6 with thermal (slow) neutrons. It beta **decays** like: T -> He3 + e

T fuses with D at a **temperature** an order of mag lower than for D-D fusion, hence its usefulness in a weapon.

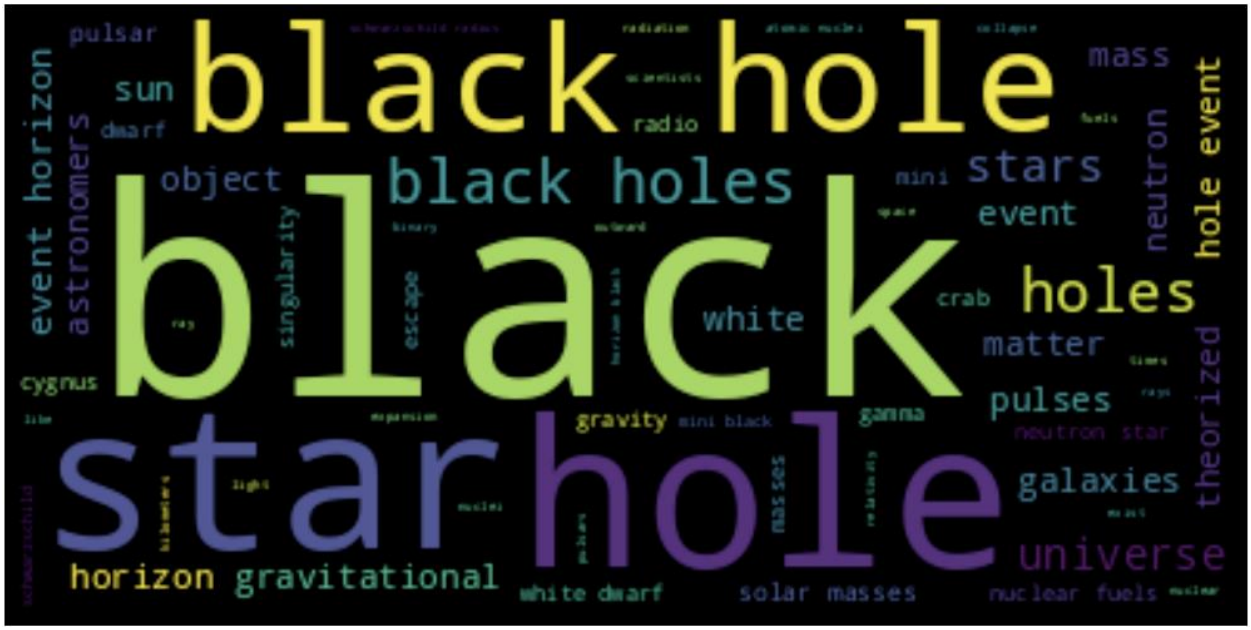
Lithium

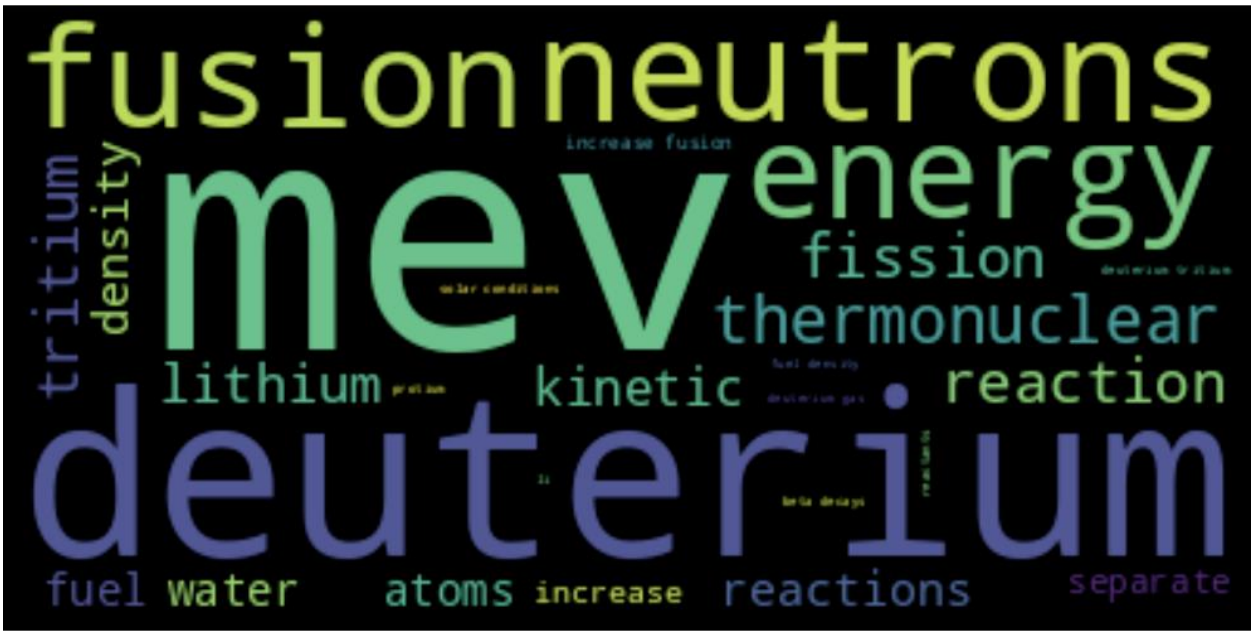
The lightest of metals, only 1/2 as dense as water. Found combined with other elements in igneous rocks and mineral spring water. Li7 is separated

Task 2: Word cloud

第二個任務是用 word cloud 視覺化檔案中最重要的關鍵字，可參考：here 中的方法（可用 pip 安裝 installed with pip）。使用提供的檔案和 tf.idf 資料，創立一個 dictionary，把最重要的關鍵字和標準化的頻率

相關聯，而後傳遞給 WordCloud 的 generate_from_frequencies()方法。具體來說，程式上會是用檔案名稱，視覺化成一個 word cloud，以及呈現最多數量關鍵字們。以下示意圖是和前一個任務使用相同檔案的結果。



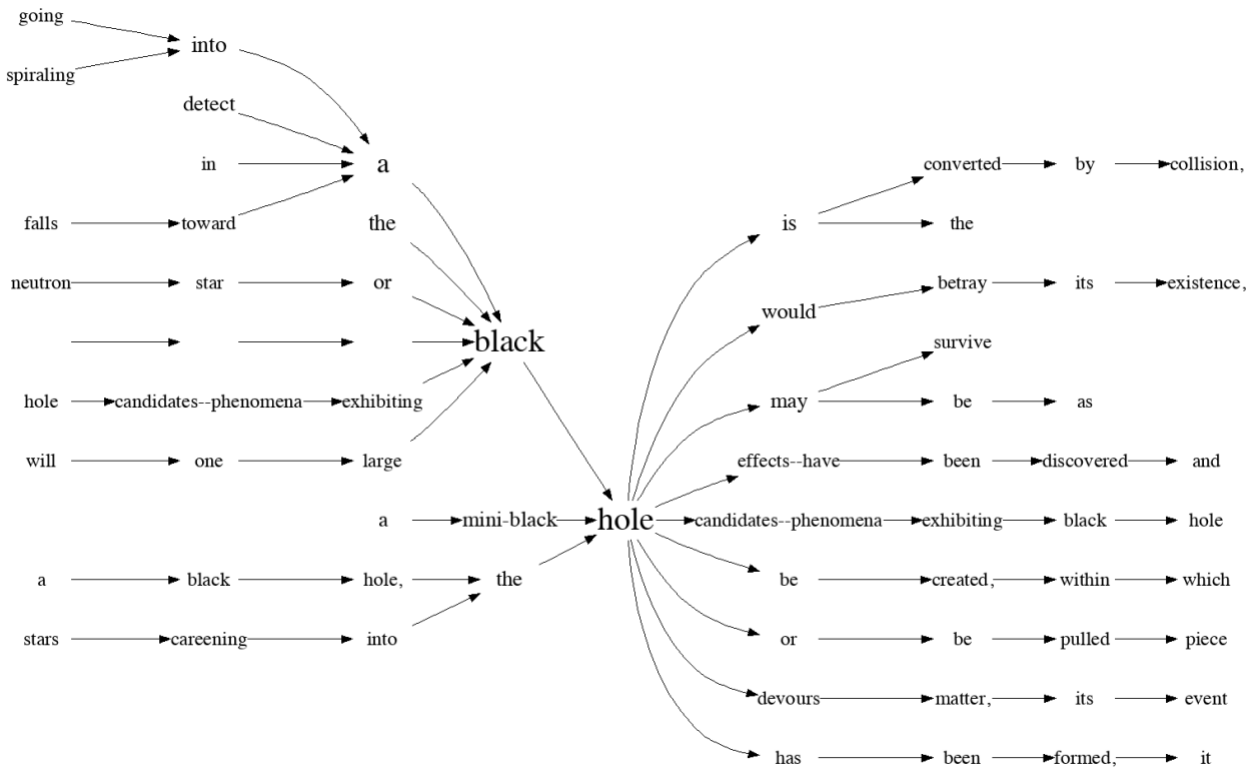


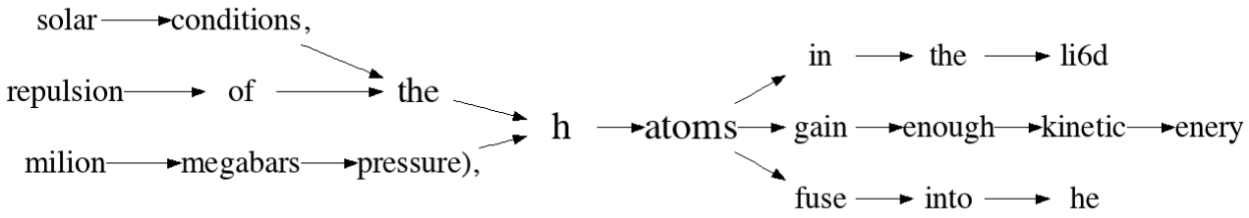
API: `python a4_wordcloud.py -d [--d] <name_of_document> -n [--number] <max_number_of_words>`

Task 3: Word tree

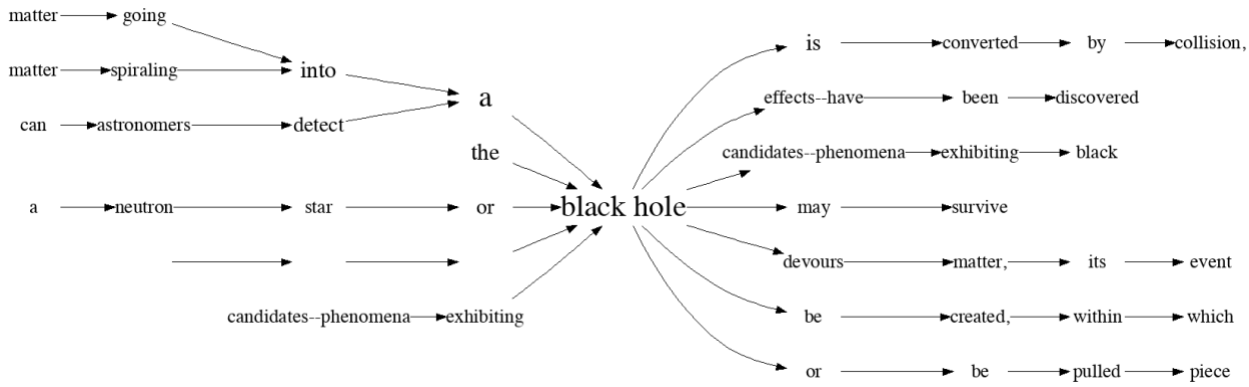
最後一個任務是使用 word tree 視覺化句子中的特定關鍵字，有幾個選項可以達成這個目標。以下的示意圖是使用這裡的方法（here），這個方法可以透 pip 安裝（pip）。在這個特定的 library 裡，文檔必須以句子的形式提供，而後透過演算法辨識包含關鍵字的部分，透過 nltk library 的 punkt sentence tokenizer (nltk.tokenize.sent_tokenize())，可以把文檔分解成句子。在程式上，需要把要處理的檔案名稱、作為 tree root 的關鍵字，還有分析句子中最大數量的詞語作為 input 的參數。有了這些資訊，城市可以創造一個 word tree 然後匯出成圖檔（wordtree library 將相應的圖像命名為"<keyword>.gv.png"，其中的<keyword>是選定的關鍵字）。

以下的示意圖是使用跟前面相同的檔案，以"hole"跟"atoms"為關鍵字的結果。





這兩個示意有修改了 wordtree library，以允許呈現由多個單詞組成的關鍵字做為 root 的結果。以 blackhol.txt 為例，使用”black hole”而不是“hole“會產生如下所示的結果。 __init__.py 修改版請見附檔 **__init__.py**。



API: `python a4_wordtree.py -d [--d] <name_of_document> -k [--keyword] <keyword(s)> -n [--number] <max_number_of_words_in_sentences>`

Data Set

science.tgz: 科學主題相關的文字檔案語料庫

tfidf.csv: 所有文件中的重要關鍵字的 tf.idf 值

__init__.py: __init__.py 修改版