Running Word2Vec with Chinese/English Wikipedia Dump

LUO Linkai

Deep Learning Research & Application Center

March 6, 2017

Outline

Word2Vec Revisited

English Wikipedia Word2Vec

Chinese Wikipedia Word2Vec

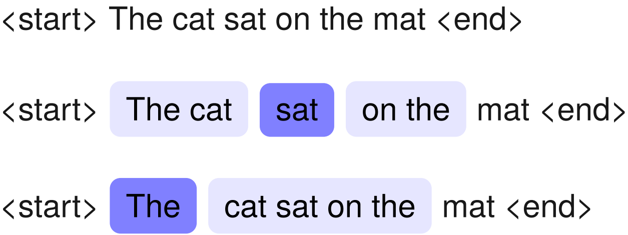
Concluding Remarks

Word2Vec Revisited

What is Word2Vec

- The corpus: “The cat sat on the mat”
- One-hot representation: “sat” = $[0, 0, 0, 1]^T$
- Distributed representation: “sat” = $[0.01, -0.02, 0.03]^T$
- Word similarity: “sat” $\approx$ “sit”
- Word2Vec is the technique to obtain such meaningful word vectors given corpus

Idea

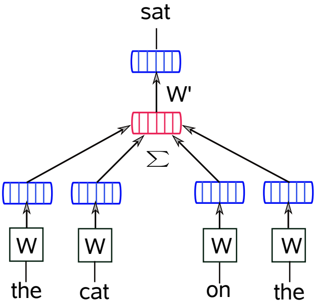

<start> The cat sat on the mat <end>

<start> The cat sat on the mat <end>

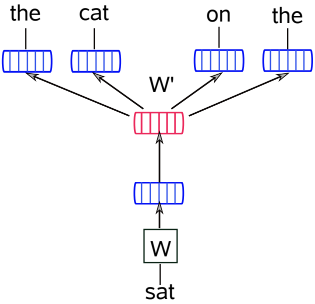
<start> The cat sat on the mat <end>

- CBOW: $p(sat|[the, cat, on, the])$ (treats an entire context as one observation)
- Skip-gram: $p(the|sat)\ p(cat|sat)\ p(on|sat)\ p(the|sat)$ (treats each context-target pair as a new observation)

Word2Vec Algorithms

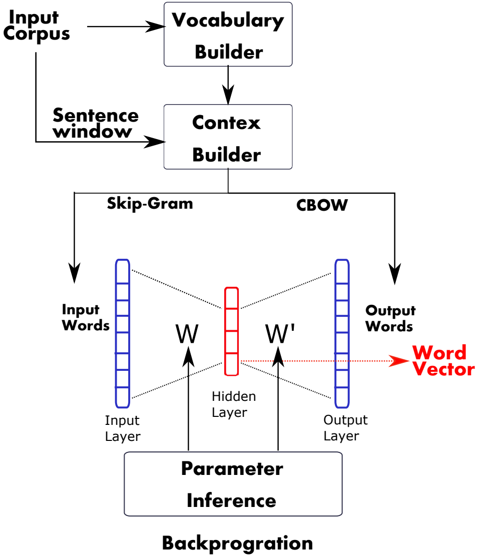


CBOW

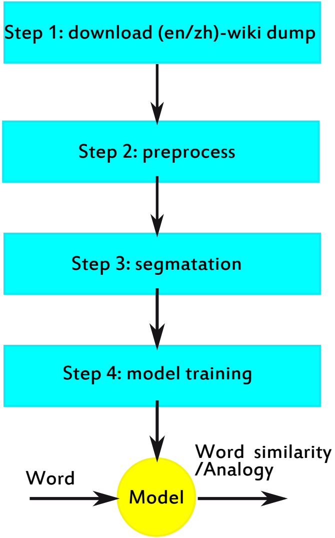


Skip-gram

Word2Vec Decomposed



Steps of Wikipedia Word2Vec



Step 1: download (en/zh)-wiki dump

Step 2: preprocess

Step 3: segmatation

Step 4: model training

English Wikipedia Word2Vec

Download English Wikipedia Dump

- `https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2`
- 13+ G file size, 4,000,000+ articles
- Preprocess, 2G+ memory required

```
# Remove words occuring less than 20 times, and words occuring in more than 10% of the documents. (keep_n is the vocabulary size)
wiki.dictionary.filter_extremes(no_below=20, no_above=0.1,^^Ikeep_n=100000)
```

Train Word2Vec Model

- Use gensim with the following settings:

```
sentences = SentencesIterator(wiki)
model = gensim.models.Word2Vec(sentences=sentences, size=200, min_count=3, window=5, workers=cores)
```

- Required RAM: $\mathcal{O}(\mathrm{size} \cdot |V|)$
- $\sim$ 24 hr training on virtual server

Word2Vec Result

- Word similarity:

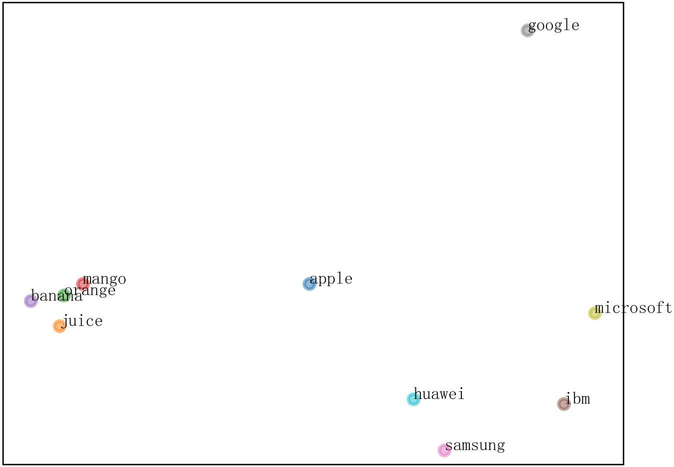
```
model = gensim.models.Word2Vec.load(os.path.join(MODEL_PATH, 'word2vec.model'), mmap='r')  # load large array
print(model.similarity('apple', 'mango'), model.similarity('apple', 'ibm'))
0.465574139702 0.531822866913
```

- Word Analogy:

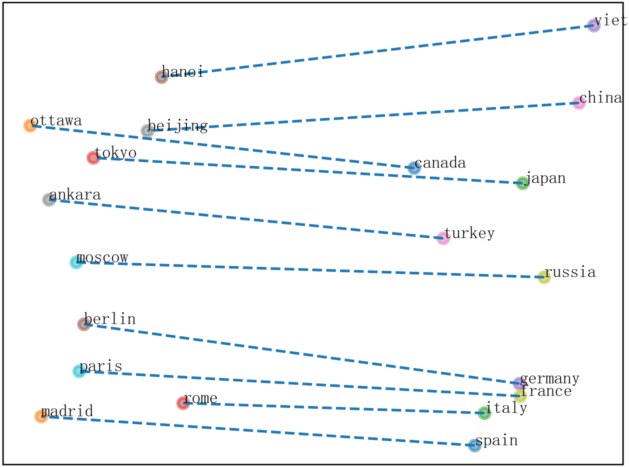
```
King–man+woman

”queen” – similarity: 0.719937
”monarch” – similarity: 0.610968
”princess” – similarity: 0.608749
”prince” – similarity: 0.607932
”empress” – similarity: 0.586417
”throne” – similarity: 0.568553
”emperor” – similarity: 0.553523
”isabeau” – similarity: 0.540711
”amalasuntha” – similarity: 0.537015
```

Word Vector: “Apple”



Word Analogy: Country-Captain Paris



Chinese Wikipedia Word2Vec

Download Chinese Wikipedia Dump

- `https://dumps.wikimedia.org/zhwiki/latest/zhwiki-latest-pages-articles.xml.bz2`
- it contains traditional Chinese and simplified Chinese articles
- 1.3+ G file size, 230,000+ articles
- Preprocess, 2G+ memory required

Preprocessing

- Use OpenCC to translate from simplified Chinese to traditional Chinese
- Developed on C++, supporting Python
-

```
openCC = OpenCC('s2twp')  # convert from Simplified Chinese to Traditional Chinese
converted = openCC.convert(original)
print(openCC.convert('这是简体中文'))
這是簡體中文
print(openCC.convert('智能手机'))
智慧手機
print(openCC.convert('一条短信'))
一條簡訊
```

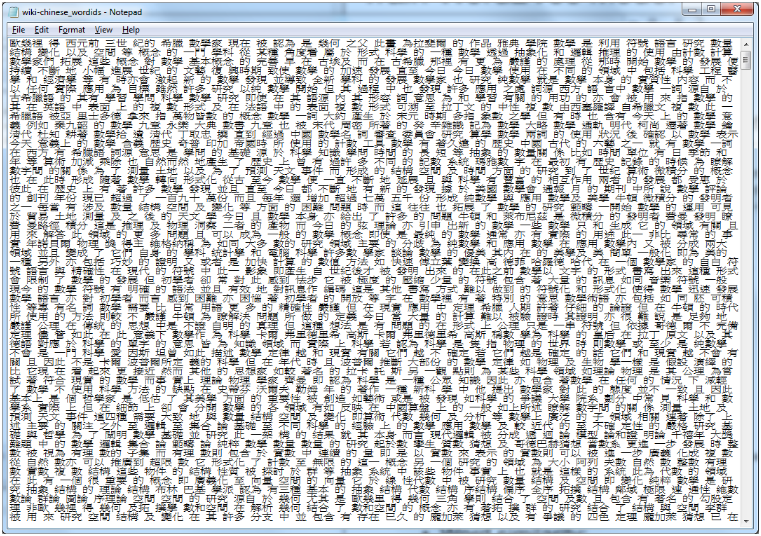
Segmentation

- English uses whitespace, dots, etc to separate words, but Chinese does not
- 下雨天留客天留我不留
    - 下雨天留客, 天留我不留
    - 下雨天, 留客天, 留我不? 留
    - ...
- Many existing tools: stanford, fudan, jieba ...

Segmentation

```
import jieba
print('/'.join(jieba.cut('下雨天留客天留我不留',cut_all=False)))
下雨天/留客/天留/我/不留
'我畢業於香港科技大學化工系'
我/畢業/於/香港科技大學/化工系
'見了他,她變得很低很低,低到塵埃里,但她心里是歡喜的,從塵埃里開出花來'
見/了/他/,/她/變得/很/低/很/低/,/低到/塵埃/里/,/但/她/心里/是/歡喜/的/,/從/塵埃/里/開出/花來
```

File Format



```
歐幾裡 得 西元前 三世 紀的 希臘 數學家 現在 被 認為 是 幾何 之父 此書 為拉斐爾 的 作品 雅典 學院 數學 是 利用 符號 語言 研究 數量
結構 變化 以及 空間 等 概念 的 一門 學科 從 某種 角度看 屬 於 形式 科學 的 一種 數學 透過 抽象化 和 邏輯 推理 的 使用 由計數 計算
數學家們 拓展 這些 概念 對 數學 基本概念 的 完善 早 在 古埃及 而 在 古希臘 那裡 有 更 為 嚴謹 的 處理 從 那時 開始 數學 的 發展 便
持續 不斷 地 小幅 進展 世紀 的 文藝 復 興時期 致使 數學 的 加速 發展 直至 今日 今日 數學 使用 在 不同 的 領域 中 包括 科學 工程 醫
學 和 經濟學 等 有 時亦會 激起 新 的 數學 發現 並導致 全新 學科 的 發展 數學家 也 研究 純數學 就是 數學 本身 的 實質性 內容 而 不
以 任何 實際 應用 為 目標 雖然 許多 研究 以純 數學 開始 但 其 過程 中 也 發現 許多 應用 之處 詞源 西方 語 言中 數學 一詞 源自 於
古希臘語 的 其有 學習 學問 科學 數學 研究 即使 在 其 語源 內 其 形容 詞 意思 為 和 學習 有關 的 用功 的 亦 會 被 用 來 指 數學 的
其 在 英語 中 表面 上 的 複 數 形式 及 在 法語 中 的 表面 複 數 形式 可溯 至 拉丁文 的 中性 複 數 由西塞羅譯 自希臘文 複 數 此 一
希臘語 被亞 里士多德 拿來 指 萬物皆數 的 概念 數學 一詞 大約 產生 於 宋元 時期 多指 象數 之學 但 有 時 也 含有 今天 上 的 數學 意
義 例如 秦九韶 的 數學 九章 永樂 大典 數書 九章 也 被 宋代 周密 所著 的 癸 辛雜識 記為 數學 大略 數學 通軌 明代 柯尚 遷著 數學 鑰
清代 杜知 耕著 數學拾 遺 清代 丁取忠 撰 直到 經過 中國 數學名 詞 審查 委員會 研究 算學 數學 兩詞 的 使用 狀況 後 確認 以 數學 表示
今天 意義上 的 數學 含義 歷史 奇普 印加 帝國時 所 使用 的 計數 工具 數學 有 著久遠 的 歷史 中國 古代 的 六藝 之一 就 有 數學 一詞
在 西方 有 希臘語 詞源 意思 是 學問 的 基礎 源 於 科學 知識 學問 時間 的 長 短 等 抽象 的 數量關 係 比如 時間 單位 有 日 季節 和
年 等 算術 加減 乘除 也 自然而然 地產生 了 歷史 上 曾 有 過許 多 不同 的 記數 系統 瑪雅數 字 在 最初 有 歷史 記錄 的 時候 為 瞭解
數字間 的關 係 為 了 測量 土地 以及 為 了 預測 天文 事件 而 形成 的 結構 空間 及 時間 方面 的 研究 到 了 世紀 算術 微積分 的 概念
也 在 此時 形成 隨著 數學 轉向 形式化 從古 至今 數學 便 一直 不斷 地 延展 且 與 科學 有 豐富 的 相互作用 兩者 的 發展 都 受惠 於
彼此 在 歷史 上 有 著 許多 數學 發現 並且 直至 今日 都 不斷 地 有 新 的 發現 據 於 美國 數學會 通報 月 的 期刊 中所 說 數學 評論
的 創刊 年份 現已 超過 了 一百九十 萬份 而且 每年 還 增加 超過 七萬 五千 份 形成 純數學 與 應用 數學及 美學 牛頓 微積分 的 發明者
之一 每當 有 涉及 數量 結構 空間 及 變化 等 方面 的 困難 問題 時 而 這 往往 也 拓展 了 數學 的 研究 範疇 一開始 數學 的 運用 可見
於 貿易 土地 測量 及 之 後 的 天文 學 今日 且 數學 本身 亦 給出 了 許多 的 問題 牛頓 和 萊布尼茲 是 微積分 的 發明者 費曼 發明 瞭
費 曼路徑 積分 這是 推理 及 物理 洞察 二者 的 產物 而 今日 的 弦 理論 亦 引申 出新 的 數學 一些 數學 只 和 生成 它 的 領域 有關 且
用 來 解答 此 領域 的 更 多 問題 且 可以 成為 一般 的 數學 概念 即使 是 最純 的 數學 通常 亦 有 實際 的 用途 此一 非比 尋常 的 事
實 年諾貝爾 物理 獎 得主 維格納稱 為 如同 大多 數的 研究 領域 主要 的 分歧 為 純數學 和 應用 數學 在 應用 數學內 又 被 分成 兩大
領域 並且 變成 了 它們 自身 的 學科 統計學 和 電腦 科學 許多數 學家 談論 數學 的 優美 其內 在 的 美學及 美 簡單 一般化 即為 美的
一種 另外 亦 包括 巧妙 的 證明 又 或者 是 加快 計算 的 數值 方法 如 快速 傅立葉 變換 高 德菲 哈羅德 哈代 在 一個 數學家 的 自白 符
號 語言 與 精確性 在 現代 的 符號 中 此一 影象 即產生 自 世紀後才 被 發明 出來 的 在此之前 數學以 文字 的 形式 書寫 出來 這種 形式
會 限制 了 數學 的 發展 但 初學者 卻 常 對 此 感到 怯步 它 被 極度 的 壓縮 少量 的 符號 包含 著 大量 的 訊息 如同 音樂 符號 一般
現今 的 數學 符號 有 明確 的 語法 並且 有效 地 對訊息作 編碼 這是 其他 書寫 方式 難以 做到 的 符號化 和 形式化 使得 數學 迅速 發展
數學 語言 亦 對 初學者 而言 感到 困難 亦 困惱 著 初學者 的 開放 等 字 在 數學裡 有 著 特別 的 意思 數學術語 亦 包括 如 同 胚 可積
性 等專 有名 詞 數學 需要 比 日常 用語 更 多 的 精確性 嚴謹 但 在 現實 應用 中 定理 希臘 人期 許著 仔細 的 論證 但 在 牛頓 的 時代
所 使用 的 方法 則較 不 嚴謹 牛頓 為 瞭解決 問題 所 做 的 定義 今日 當 大量 的 計算 難以 被驗 證時 其證明 亦 很 難 說 是 足夠 地
嚴謹 公理 在 傳統 的 思想 中是 不證 自明 的 真理 但 這種 想法 是 有 問題 的 在 形式 上 公理 只是 一串 符號 但 依據 哥德 爾 不 完備
定理 儘 管 如此 在 此 意義下 數學作 為 科學 卡爾 弗里德里希 高斯 卡爾 弗里德里希 高斯 稱數 學為 科學 的 皇后 在 拉丁 原文 以及 其
德語 對應 於 科學 的 單字 的 意思 皆 為 知識 領域 而 實際 上 科學 若 認為 科學 是 隻 指 物理 的 世界 時 則數學 或 至少 是 純數學
不會 是 一門 科學 愛 因斯 坦曾 如此 描述 數學 定律 越 和 現實 有關 它們 越 不 確定 若 它們 越是 確定 的 話 它們 和 現實 越 不會 有
關 且 因此 不是 卡爾 波普爾所定 義的 科學 但 在 年代 時 且 波普爾 推斷 大部份 的 數學 定律 如 物理 及 生物 學一樣 是 假設 演繹 的
比 它現 在 看 起來 更 接近 然而 其他 的 思想家 如較 著名 的 拉卡 託 斯 另 一觀 點則 為 某些 科學 領域 如理論 物理 是 其 公理 為嘗
試 著 符合 現實 的 數學 而事 實上 理論 物理 學家 齊曼 即 認為 科學 是 一種 公眾 知識 因此 亦 包含 著數學 在 任何 的 情況 下 減輕
了 數學 不 使用 科學 方法 的 缺點 在 史蒂芬 沃爾夫 勒姆 年 的 著作 一種 新科 學 中 他 提出 數學家 對 此 的 態度 並不 一致 且 因此
基本上 是 個 哲學家 是 低估 了 其美學 方面 的 重要性 被 創造 如藝術 或是 被 發現 如科學 的 爭議 大學 院系 劃分 中常 見 科學 和 數
學系 實際 上 但 在 細節 上 卻 會 分開 數學 的 各 領域 有如 反映 在 中國算盤 上 的 一般 如上所述 瞭解 數字間 的關 係 測量 土地 及
預測 天文 事件 這四種 需要 大致 地 與 數量 結構 空間 及 變化 即算術 代數 幾何 及 分析 等 數學上 廣泛 的 子 領域 相關 連著 除了 上
述 主要 的 關注 之外 至 邏輯 至 集合 論 基礎 至 不同 科學 的 經驗 上 的 數學 應用 數學 及 較 近代 的 至 不確 定性 的 嚴格 研究 基
礎 與 哲學 為 了 闡明 數學 基礎 並 研究 此 一架 構 的 結果 就 其 本身 而言 現代邏輯 被 分成 遞 迴 論 模型 論和證 明論 千禧年 大獎
難題 中 的 數學 邏輯 集合 論 範疇 論 純粹 數學 數量 數量 的 研究 起於數 學生 質數 猜想 及 哥德巴赫猜想 當數系 更進 一步 發展 時 整
數 被 視為 有理 數的 子集 而 有理 數則 包含 於 實數 中 連續 的 量 即 是 以 實數 來 表示 的 實數則 可以 被 進 一步 廣義 化成 複 數
從 自然 數亦 可以 推廣到 超限 數 它 形式化 了 計數 至 無限 的 這一 概念 另 一個 研究 的 領域 為 大小 阿列 夫數 自然 數 整數 有理
數 實數 複 數 結構 這些 物件 的 結構 性質 被 探討 於 群 等 抽象 系統 中 該些 物件 事實 上 也 就是 這樣 的 系統 此為 代數 的 領域
在 此 有 一個 很 重要 的 概念 即 廣義化 至 向量 空間 的 向量 它 於 線 性代數 中 被 研究 數量 結構 及 空間 即 變化 純粹 數學 是 研
究 抽象 結構 的 理論 結構 布林 巴基 學派 認為 有三種 基本 的 抽象 結構 代數 結構 序結構 偏序 全序 拓撲 結構 鄰域 極限 連 通性 維數
數論 群論 圖論 序理論 空間 空間 的 研究 源自 於 幾何 尤其 是 歐幾里 得 幾何 三角 學則 結合 了 空間 及數 且 包含 有 著名 的 勾股定
理 非歐 幾裡 得 幾何 及拓 撲學 數和空間 在 解析 幾何 結合 了 數和空間 的 概念 亦 有 著拓 撲 群 的 研究 結合 了 結構 與 空間 李群
被 用 來 研究 空間 結構 及 變化 在 其 許多 分支 中 並 包含 有 存在 已久 的 龐加萊 猜想 以及 有 爭議 的 四色 定理 龐加萊 猜想 已 在
```

Train Chinese Word2Vec Model

- Use gensim with the following settings:

```
sentences = SentencesIterator(wiki)
model = gensim.models.Word2Vec(sentences=sentences, size=300, min_count=3, window=5, workers=cores)
```

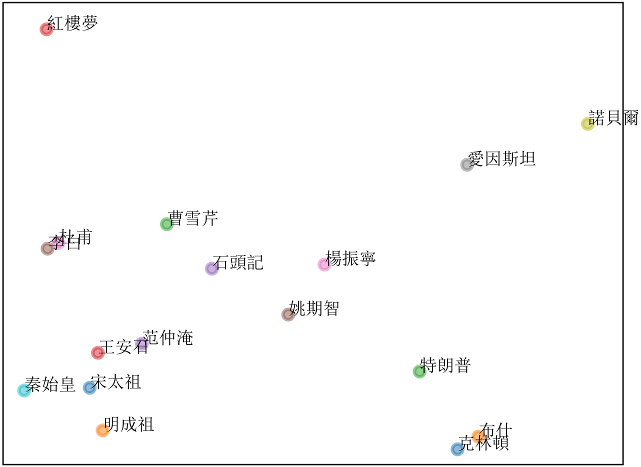
- $\sim$ 12 hr training

Chinese Word2Vec Result

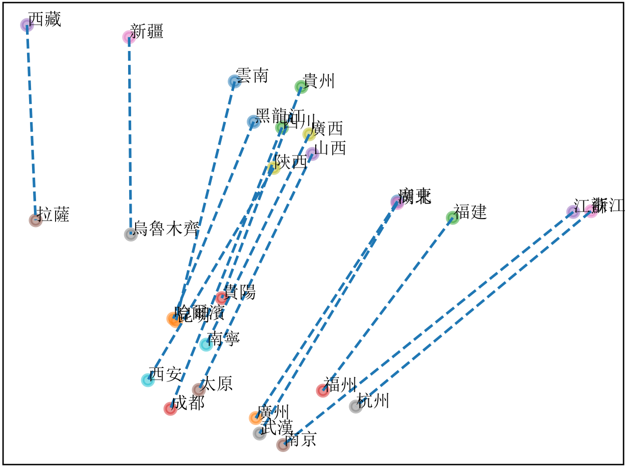
```
model = gensim.models.Word2Vec.load(os.path.join(MODEL_PATH, 'word2vec.model'), mmap='r')
print(model.similarity('蘇軾', '東坡'))
0.612037855344

pprint(model.most_similar('奧斯卡'))
[('奧斯卡金像獎', 0.6850196123123169),
('奧斯卡獎', 0.6607037782669067),
('金球獎', 0.6433079242706299),
('東尼獎', 0.579437375068664 6),
('金棕櫚獎', 0.5717018842697144),
('艾美獎', 0.5554369688034058),
('斷背山', 0.5518748760223389),
('泰坦尼克號', 0.5482996702194214),
('布克獎', 0.546217679977417),
('學院獎', 0.5445188879966736)]
```

Word Similarity: Names



Word Analogy: Province-Capital Paris



Concluding Remarks

- Running Word2Vec with English/Chinese Wikipedia dump have acceptable performance in terms of word similarity/analogy
- Judge of Word2Vec’s performance should also take machine translation into consideration
- Next …
    - sentence pairs
    - constructing basic LSTMs