

(۱)

در گام پیش پردازش مراحل زیر را طی کرده ایم:

- متون را نرمالایز کرده ایم
- Token هارا استخراج کرده ایم
- ریشه یابی کرده ایم
- کلمات پرتکرار (stop-words) هارا حذف کرده ایم

برای سه مورد اول از کتابخانه parsivar و در مورد آخر از کتابخانه hazm استفاده شده است.

که توابع آن به صورت زیر پیاده سازی شده اند:

```
def normalize(input_text):
    output_text = my_normalizer.normalize(input_text)
    return output_text

def get_tokens(input_text):
    words = my_tokenizer.tokenize_words(my_normalizer.normalize(input_text))
    return words

def get_stems(input_words):
    words = []
    for word in input_words:
        words.append(my_stemmer.convert_to_stem(word))
    return words

def delete_stop_words(input_words):
    news_words = []
    stop_words = stopwords_list()
    for word in input_words:
        if word not in stop_words:
            news_words.append(word)

    return news_words
```

فاز اول پروژه

علت نرمالایز سازی : استفاده از نرمال سازی باعث میشود تمام متون از نظر فاصله و نیم فاصله، تاریخ ها، اعداد و ... یکسان شوند و مثلا در متنی عدد ها فارسی و در متن دیگر انگلیسی نباشند و این قواعد در تمام متون به صورت یکسان رعایت شود.

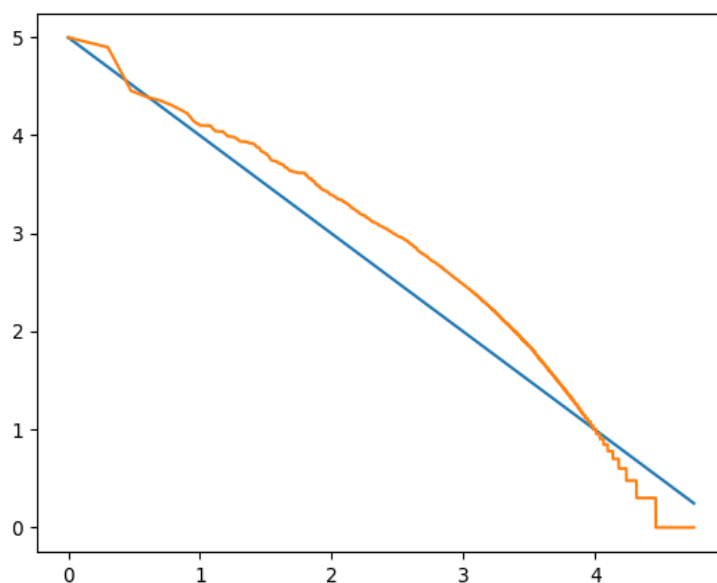
علت استخراج توکن : قسمت اصلی ساخت موتور های جست و جو ساختن توکن هاست که درواقع به وسیله توکن های میتوانیم لیست شاخص های مکانی را بسازیم و توسط آن به کوئری کاربر پاسخ دهیم.

علت ریشه یابی : ریشه یابی باعث میشود کلمه هایی مانند: رفت، رفتم، رفتی، رفتیم، رفتند، دارم میرم، داشتم میرفتم و.... و تمامی اشکال فعل ها و اسم ها تنها به یک کلمه تبدیل شود و زمانی که کاربر هرکدام از موارد ذکر شده را وارد کرد اسناد مربوط به دیگر حالت ها نیز برگردانده شوند.

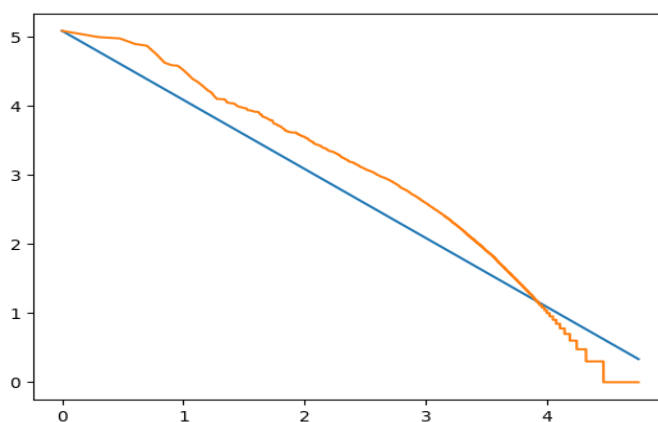
علت حذف کلمات پرتکرار : علت این کار هم این است که برای جست و جو کاربر کلماتی نظیر است، تو، "،"، کوچک و ... از لیست توکن ها حذف شوند زیرا برای کاربر وجود این کلمات در کوئری در نتایج برگردانده شده تاثیری تقریبا ندارد و کاهش حجم بسیار ارزنده تر از مقدار کمی کاهش دقت در نتایج بازگردانده شده است.

(۲)

نمودار برای قبل از حذف کردن stop-word ها به صورت زیر است:



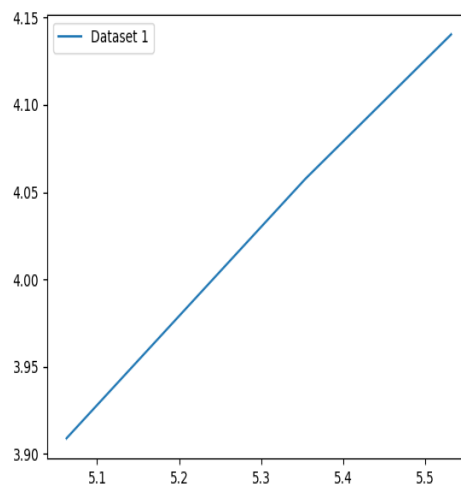
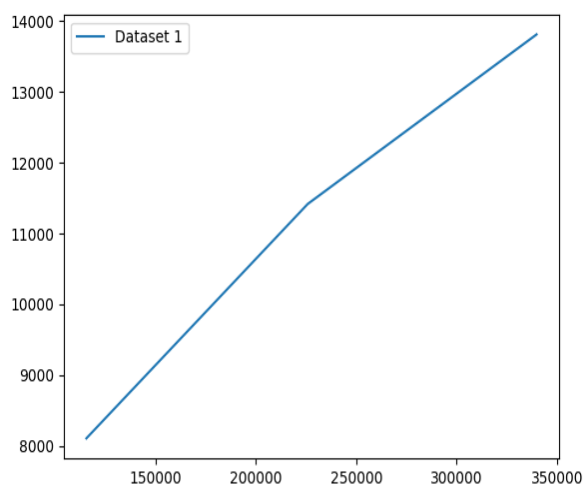
نمودار برای بعد از حذف کردن stop-word ها به صورت زیر است:



میبینیم که در دو حالت قانون برقرار است و نمودار ها تقریباً بر هم منطبق هستند.

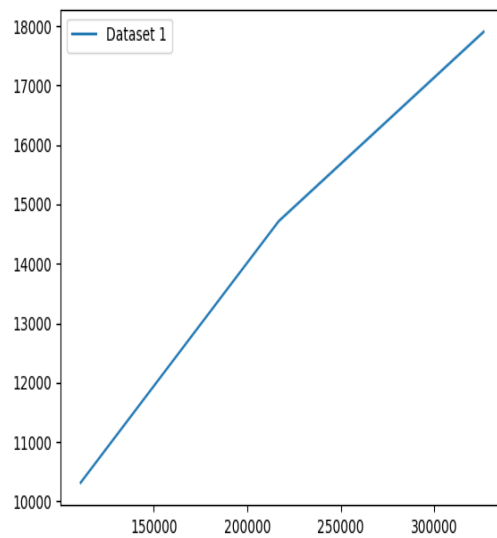
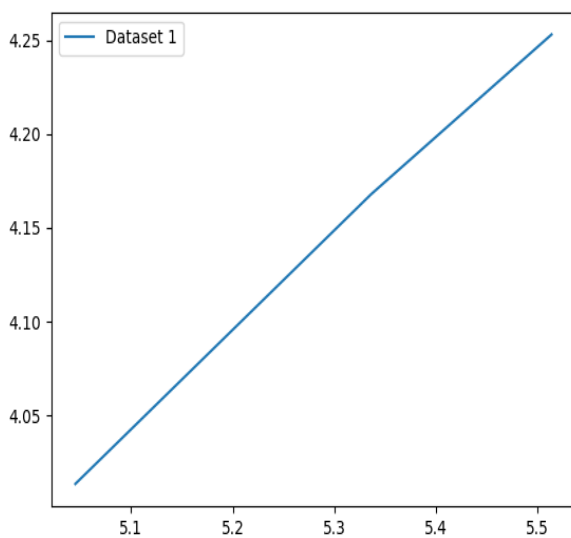
(۳

نمودار های زیر مربوط به بعد از ریشه یابی است:



که یکی از آنها لوگاریتمی و دیگری بدون لوگاریتم گیری رسم شده

نمودار های زیر مربوط به قبل از ریشه یابی است:



که یکی از آنها لوگاریتمی و دیگری بدون لوگاریتم گیری رسم شده

میبینیم که در دو حالت نمودار ها منطبق بر هم هستند.

در حالت قبل از ریشه یابی داریم:

تعداد توکن ها: ۳۲۶۳۹۲

تعداد کلمه ها: ۱۷۹۰۴

تعداد کل توکن ها: ۱۹۵۹۵۵۲

تعداد کل کلمه ها: ۷۲۸۹۰

که به ازای $b = 0.67$ رابطه برقرار می شود.

در حالت بعد از ریشه یابی داریم:

تعداد توکن ها: ۳۳۹۸۸۳

تعداد کلمه ها: ۱۳۸۱۲

تعداد کل توکن ها: ۲۰۷۸۶۸۴

تعداد کل کلمه ها: ۵۵۸۹۰

که به ازای $b = 0.66$ رابطه برقرار می شود.

(۴)

- کلمه دادند که بعد از ریشه یابی به کلمه "ده" تبدیل شد که میتواند با عدد ده اشتباه گرفته شود
- کلمه است که بعد از ریشه یابی به کلمه "اس" تبدیل شده است و لزومی برای ریشه یابی وجود نداشت
- کلمه می افتد که به معنای اتفاق افتادن هست بعد از ریشه یابی به کلمه "افت" تبدیل شده است که در اسناد ورزشی ممکن است به معنای افت تیم باشد و بهتر بود از ریشه افتاد به جای افت استفاده میکردیم

(۵

(الف

```

C:\Python38\python.exe "D:/College/Term 7/Informations Retrieval/Project/Phase 1/search_words.py"
{1: [153, 410, 733, 767, 976, 994, 1008, 1073, 1331, 1577, 1729, 1743, 1753, 1769, 1772, 1787, 1830, 1831, 1847, 1898, 1938, 1944, 2025, 2030, 2037, 2041, 2110, 2131, 2181, 2217, 2263, 2277, 2283, 2296]}
doc_id: 153
title: توضیحات مسؤول مسابقات لیگ یک درباره شایعه سکه ناظر بازی
=====
doc_id: 410
title: اعلام آخرین اقدامات تراکتور برای بازشدن پنجره/ احتمال استفاده از بازیکنان جدید برای گل‌گهر
=====
doc_id: 733
title: گزارش تمرین پرسپولیس | روحیه شاد قبل از مصاف با الهلال/ پا به توپ شدن گل‌محمدی و باقری
=====
doc_id: 767
title: واکنش نصیرزاده به اظهارات مالک ماشین سازی: به جای فرافکنی به تعهداتان عمل کنید
=====
doc_id: 976
title: باج بن سلمان به انگلیس و قطر/ پروژه فوتبالی آل سعود چگونه رقم خورد؟
=====
doc_id: 994
title: پنجره نقل و انتقالات باشگاه فولاد خوزستان باز شد
=====
doc_id: 1008
title: معاون بین‌الملل و مدیر کمیته حرفه ای سازی باشگاه استقلال منصوب شدند
=====
doc_id: 1073
title: برای هتل محل اقامت AFC باشگاه پرسپولیس خواستار ارسال تأییدیه
=====
doc_id: 1331
title: تأیید حضور تماشاگر در دیدار تیم ملی فوتبال مقابل کره جنوبی توسط وزیر ورزش
=====
doc_id: 1577
title: نامه پرسپولیس به کنفدراسیون فوتبال آسیا برای اسکان شاگردان گل محمدی در ریاض

```

```
C:\Python38\python.exe "D:/College/Informations Retrieval/Project/Phase 1/search_words.py"
{1: [1753, 1959, 2131, 2792, 2793, 7236], 2: [362, 390, 393, 552, 560, 724, 775, 978, 1807, 1146, 1228, 1249, 1702, 1704, 1710, 1716, 1721, 1724, 1725, 1728, 1737, 1751, 1753, 1759, 1760, 1765, 1778, 1780, 1781, 1782, 1783, 1784, 1785, 1786, 1787, 1788, 1789, 1790, 1791, 1792, 1793, 1794, 1795, 1796, 1797, 1798, 1799, 1800, 1801, 1802, 1803, 1804, 1805, 1806, 1807, 1808, 1809, 1810, 1811, 1812, 1813, 1814, 1815, 1816, 1817, 1818, 1819, 1820, 1821, 1822, 1823, 1824, 1825, 1826, 1827, 1828, 1829, 1830, 1831, 1832, 1833, 1834, 1835, 1836, 1837, 1838, 1839, 1840, 1841, 1842, 1843, 1844, 1845, 1846, 1847, 1848, 1849, 1850, 1851, 1852, 1853, 1854, 1855, 1856, 1857, 1858, 1859, 1860, 1861, 1862, 1863, 1864, 1865, 1866, 1867, 1868, 1869, 1870, 1871, 1872, 1873, 1874, 1875, 1876, 1877, 1878, 1879, 1880, 1881, 1882, 1883, 1884, 1885, 1886, 1887, 1888, 1889, 1890, 1891, 1892, 1893, 1894, 1895, 1896, 1897, 1898, 1899, 1900, 1901, 1902, 1903, 1904, 1905, 1906, 1907, 1908, 1909, 1910, 1911, 1912, 1913, 1914, 1915, 1916, 1917, 1918, 1919, 1920, 1921, 1922, 1923, 1924, 1925, 1926, 1927, 1928, 1929, 1930, 1931, 1932, 1933, 1934, 1935, 1936, 1937, 1938, 1939, 1940, 1941, 1942, 1943, 1944, 1945, 1946, 1947, 1948, 1949, 1950, 1951, 1952, 1953, 1954, 1955, 1956, 1957, 1958, 1959, 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026, 2027, 2028, 2029, 2030, 2031, 2032, 2033, 2034, 2035, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2045, 2046, 2047, 2048, 2049, 2050, 2051, 2052, 2053, 2054, 2055, 2056, 2057, 2058, 2059, 2060, 2061, 2062, 2063, 2064, 2065, 2066, 2067, 2068, 2069, 2070, 2071, 2072, 2073, 2074, 2075, 2076, 2077, 2078, 2079, 2080, 2081, 2082, 2083, 2084, 2085, 2086, 2087, 2088, 2089, 2090, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2100, 2101, 2102, 2103, 2104, 2105, 2106, 2107, 2108, 2109, 2110, 2111, 2112, 2113, 2114, 2115, 2116, 2117, 2118, 2119, 2120, 2121, 2122, 2123, 2124, 2125, 2126, 2127, 2128, 2129, 2130, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2143, 2144, 2145, 2146, 2147, 2148, 2149, 2150, 2151, 2152, 2153, 2154, 2155, 2156, 2157, 2158, 2159, 2160, 2161, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2176, 2177, 2178, 2179, 2180, 2181, 2182, 2183, 2184, 2185, 2186, 2187, 2188, 2189, 2190, 2191, 2192, 2193, 2194, 2195, 2196, 2197, 2198, 2199, 2200, 2201, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 2221, 2222, 2223, 2224, 2225, 2226, 2227, 2228, 2229, 2230, 2231, 2232, 2233, 2234, 2235, 2236, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 2248, 2249, 2250, 2251, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266, 2267, 2268, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277, 2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288, 2289, 2290, 2291, 2292, 2293, 2294, 2295, 2296, 2297, 2298, 2299, 2300, 2301, 2302, 2303, 2304, 2305, 2306, 2307, 2308, 2309, 2310, 2311, 2312, 2313, 2314, 2315, 2316, 2317, 2318, 2319, 2320, 2321, 2322, 2323, 2324, 2325, 2326, 2327, 2328, 2329, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2337, 2338, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2347, 2348, 2349, 2350, 2351, 2352, 2353, 2354, 2355, 2356, 2357, 2358, 2359, 2360, 2361, 2362, 2363, 2364, 2365, 2366, 2367, 2368, 2369, 2370, 2371, 2372, 2373, 2374, 2375, 2376, 2377, 2378, 2379, 2380, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2390, 2391, 2392, 2393, 2394, 2395, 2396, 2397, 2398, 2399, 2400, 2401, 2402, 2403, 2404, 2405, 2406, 2407, 2408, 2409, 2410, 2411, 2412, 2413, 2414, 2415, 2416, 2417, 2418, 2419, 2420, 2421, 2422, 2423, 2424, 
```



```
C:\Python38\python.exe "D:/College/Term 7/Informations Retrieval/Project/Phase 1/search_words.py"
{1: [1973, 2025, 2298, 2301, 2398, 2426, 2726, 2741, 2743, 2751, 2770, 2781, 2785, 2861, 2903, 3374, 4767, 4772, 5086, 6567, 6799, 6873, 6911, 7543], 2: [1749, 1772, 1802, 1972, 1973, 2025, 2038, 2053,
doc_id: 1973
title: گزارش نظارت میدانی نمایندگان از مرزهای شمال‌غرب به کمیسیون امنیت ملی
=====
doc_id: 2025
title: محسن اسلامی مدیرکل دفتر امور سیاسی وزارت کشور شد
=====
doc_id: 2298
title: اعتراض هیات پارلمانی ایران به نماینده منصور هادی در اجلاس تغییرات آب و هوایی/ همکاری دولت مستعفی یمن با ائتلاف سعودی علیه یمنی‌ها
=====
doc_id: 2301
title: هیئت پارلمانی ایران به رم سفر کرد
=====
doc_id: 2398
title: نقض حاکمیت ملی عراق» با قراردادی 50ساله میان اقلیم و ترکیه»
=====
doc_id: 2426
title: راهبرد دولت آیت‌الله رئیسی برای شکست تحریم‌ها ۳
=====
doc_id: 2726
title: جان‌مایه سخنان رئیس‌جمهور
=====
doc_id: 2741
title: آصفی: اظهارات رئیسی در سازمان ملل اقدامی برای رسوا کردن غربی‌ها و احقاق حقوق مردم بود
=====
doc_id: 2743
title: راستینه: اظهارات آیت الله رئیسی مطالبه‌گری انقلابی در سازمان ملل بود/ تاکید رئیسی جمهور بر منافع ملت
=====
doc_id: 2751
title: منتظر منافع عضویت ایران در سازمان شانگهای با 40 درصد مساحت دنیا باشید/ توسعه همکاری‌های حمل و نقل، انرژی و بانکی
=====
```

```
C:\Python38\python.exe "D:/College/Term 7/Informations Retrieval/Project/Phase 1/search_words.py"
{1: [2793, 5021, 5022], 2: [362, 1725, 1753, 1779, 1959, 2091, 2131, 2447, 2588, 2631, 2709, 2791, 2793, 3031, 4854, 5021, 5022, 5154, 7043], 3: [311, 362, 390, 393, 480, 521, 535, 552, 560, 566, 724,
doc_id: 2793
title: باید برای ثبت نقشی دانشگاهیان در دوران دفاع مقدس کار تحقیقاتی صورت گیرد
=====
doc_id: 5021
title: دفترچه راهنمای آزمون استخدامی دانشگاه‌ها برای بار چهارم اصلاح شد/تمدید مجدد مهلت ثبت‌نام
=====
doc_id: 5022
title: دفترچه راهنمای آزمون استخدامی دانشگاه‌ها برای بار چهارم اصلاح شد/تمدید مجدد مهلت ثبت‌نام
=====
doc_id: 362
title: دایی: می‌خواهم مردم مرا به عنوان انسان به یاد بیاورند نه دایی
=====
doc_id: 1725
title: سیدمحسن دهنوی عضو هیئت امنای صندوق نوآوری و شکوفایی شد
=====
doc_id: 1753
title: نامه جمعی از اساتید و متخصصان/ آقای رئیس‌جمهور در گام دوم انقلاب به داد «مدیریت» در کشور برسید
=====
doc_id: 1779
title: وزیر علوم: علم و عقل دو بال دانایی است/ علم باید برای جامعه ثروت‌آفرین باشد
=====
doc_id: 1959
title: نامه ۸ بسیج دانشجویی دانشگاه‌های تهران به معاون اول رئیس‌جمهور
=====
doc_id: 2091
title: حجت‌الاسلام رستمی فقدان فعال دانشجویی دانشگاه شریف را تسلیت گفت
=====
doc_id: 2131
title: برگداشت شهدای مسجد قدور در مقابل کنسولگری افغانستان/ آمریکا و آل سعود مقصران اصلی جنایت در افغانستان
=====
```

```
C:\Python38\python.exe "D:/College/Term 7/Informations Retrieval/Project/Phase 1/search_words.py"
{1: [80, 314, 617, 624, 844, 1331, 1601, 1704, 1706, 1708, 1732, 1743, 1748, 1749, 1753, 1763, 1769, 1776, 1787, 1798, 1803, 1804, 1809, 1812, 1820, 1837, 1842, 1854, 1906, 1909, 1910, 1913, 1926, 1928,
doc_id: 80
title: و اجرای قرارداد با شرکت اسرائیلی VAR واکنش تند فدراسیون فوتبال به حواشی
=====
doc_id: 314
title: جوابیه باشگاه آلومینیوم نسبت به محرومیت 2 بازیکن ملی‌پوش
=====
doc_id: 617
title: آمادگی قوه قضائیه برای حمایت حقوقی و معنوی از ورزشکاران/ اژه‌ای: به دعاوی حقوقی ورزشی به صورت تخصصی رسیدگی خواهد شد
=====
doc_id: 624
title: اسکی صاحب کرسی جهانی شد
=====
doc_id: 844
title: بیانیه فدراسیون تنیس درباره عدم صدور ویزای سه ملی پوش برای حضور در دیویس کاپ بحرین
=====
doc_id: 1331
title: تأیید حضور تماشاگر در دیدار تیم ملی فوتبال مقابل کره جنوبی توسط وزیر ورزش
=====
doc_id: 1601
title: تمجید فدراسیون جهانی هندبال از تاریخ‌سازی زنان ایرانی
=====
doc_id: 1704
title: سرلشکر باقری: با وجود تمام فشارهای دشمن، جمهوری اسلامی روز به روز قوی‌تر شده است
=====
doc_id: 1706
title: رئیس‌جمهور: اتحادیه اروپایی نباید تحت تأثیر فشارها و یکجانبه‌گرایی آمریکا قرار گیرد
=====
doc_id: 1708
title: اروپا جوری عمل نکند که آمریکا، اروپا را حوزه نفوذ خود بداند
=====
```

(ت)

```
C:\Python38\python.exe "D:/College/Term 7/Informations Retrieval/Project/Phase 1/search_words.py"
{1: [632, 1367, 3615, 3664, 3878, 4056, 4188]}
doc_id: 632
title: خیرخواه: برخی به دنبال فلج کردن ژمناسستیک هستند/ با بایکوت فدراسیون موفقیت‌ها بیشتر شد
=====
doc_id: 1367
title: هشدار هیات ژمناسستیک تهران در خصوص سالن‌های مختلط و اقدامات غیراخلاقی
=====
doc_id: 3615
title: دبیر مجمع فدراسیون ژمناسستیک مشخص شد
=====
doc_id: 3664
title: ثبت نام ۱۳ نامزد برای پست ریاست فدراسیون ژمناسستیک + اسامی
=====
doc_id: 3878
title: جزییات تعطیلی ورزش ایران تا پایان تیرماه+ تصویر
=====
doc_id: 4056
title: !دبیر: اگر من در مباحث فنی ۱۰ باشم، درستکار ۱۰۰ است/ بنا کاملاً بر اساس چرخه انتخابی عمل کرد
=====
doc_id: 4188
title: جزییات تعطیلی‌های ورزش ایران تا ۹ مهر ۱۴۰۰/ کدام فعالیت‌های ورزشی در تهران ممنوع است؟
=====

Process finished with exit code 0
|
```

(ث)

```
C:\Python38\python.exe "D:/College/Term 7/Informations Retrieval/Project/Phase 1/search_words.py"
{1: [4931, 5569, 5685, 5823, 5825, 5831, 5833, 5857, 6336], 2: [327, 374, 640, 782, 1021, 1032, 1034, 1075, 1083, 1106, 1146, 1160, 1337, 1378, 1445, 1556, 1579, 1703, 1737, 1752, 1754, 1759, 1767, 1781]}
doc_id: 4931
title: محموله ۱.۴ میلیون دوری واکسن کرونا وارد کشور شد
=====
doc_id: 5569
title: محموله ۱.۴ میلیون دوری واکسن کرونا وارد کشور شد
=====
doc_id: 5685
title: مهمترین سلاح مبارزه با کرونا
=====
doc_id: 5823
title: واکسناسیون؛عقلانی‌ترین راه مقابله با کرونا/پرهیز از تزریق واکسن خارج از چرخه عمومی واکسناسیون
=====
doc_id: 5825
title: نکاتی که باید در مورد واکسناسیون کرونا بدانیم
=====
doc_id: 5831
title: واکسناسیون؛عقلانی‌ترین راه مقابله با کرونا/پرهیز از تزریق واکسن خارج از چرخه عمومی واکسناسیون
=====
doc_id: 5833
title: نکاتی که باید در مورد واکسناسیون کرونا بدانیم
=====
doc_id: 5857
title: واکسن‌های کرونا با چه داروهایی تداخل دارند؟
=====
doc_id: 6336
title: امکان ایجاد لخته خون در واکسن آسترانکا چقدر است؟
=====
doc_id: 327
title: تارتار: 3 امتیاز بالارزش در آبدان بدست آوردیم/نفت این فصل از سال گذشته بهتر است
=====
```