

تمرین دوم درس روش‌های آماری در پردازش زبان طبیعی

سارا شاه‌محمدی

۸۳۰۵۹۶۰۱۹

سوال ۱-

در حوزه سرقت علمی و ادبی و جستجوی literature، ابزارهای بسیار زیادی وجود دارد. در زیر، ابتدا به معرفی نرم‌افزار HelioBlast و روش‌های پیاده‌سازی آن می‌پردازیم و در ادامه، معرفی کوتاهی از ابزارهای دیگر شباهت‌یابی متنی خواهیم آورد.

۱- سرویس شباهت‌یابی متنی **HelioBlast**¹ (با نام قدیمی etBlast) رکوردهای متنی مشابه متن ورودی را به عنوان خروجی، همراه با درصد شباهت (که در تشخیص سرقت علمی و ادبی مورد استفاده قرار می‌گیرد)، تحویل می‌دهد. با قرار دادن تنظیماتی برای این نرم‌افزار می‌توان مثلاً دیتابیس خاصی را جستجو کرد یا آن را متناسب با کاربرد خود کرد.

در نسخه وب این نرم‌افزار، متن ورودی دریافت می‌شود و با عناوین و چکیده‌های Medline/PubMed مقایسه می‌شوند. طول متن ورودی باید حداکثر ۱۰۰۰ کلمه باشد؛ به همین علت نرم‌افزار پیشنهاد کرده که چکیده/پاراگراف جستجو کنیم.²

بزرگترین انگیزه ساخت HelioBlast تهیه ابزاری بود که دانشجویان پزشکی به کمک آن بتوانند به راحتی جستجو کنند و به کارهای پیشین انجام‌شده در یک حوزه خاص دسترسی آسان داشته باشند. در زیر خلاصه‌ای از روش‌های استفاده‌شده برای پیاده‌سازی HelioBlast به اختصار آورده شده‌اند که گزارش خلاصه‌ای از مقاله تیم پیاده‌سازی این نرم‌افزار، لوپیس و همکاران (۲۰۰۶)³ است.

¹ <https://helioblast.heliotext.com>

² همان

³ James Lewis, Stephan Ossowski, Justin Hicks, Mounir Errami, Harold R. Garner, Text similarity: an alternative way to search MEDLINE, *Bioinformatics*, Volume 22, Issue 18, 15 September 2006, Pages 2298–2304

به گفته لوییس و همکاران، الگوریتم‌های هم‌ترازسازی⁴ برای شباهت‌یابی در سندهایی حتی کمی طولانی کارآمد نیستند و بسیار کند عمل می‌کنند. به همین علت، پیاده‌سازی HelioBlast در دو مرحله انجام شده است: الف) در مرحله اول، از یک الگوریتم جستجوی شباهت متنی⁵ استفاده می‌شود و ۴۰۰ سندی که بیشترین شباهت را با متن ورودی دارند، شناسایی می‌شوند. ب) در مرحله بعد، این ۴۰۰ سند با استفاده از روش‌های هم‌ترازسازی دوباره رتبه‌بندی می‌شوند.

لوییس و همکاران می‌گویند در فرایند پیاده‌سازی و ساخت HelioBlast، در الگوریتم‌های TSS شان از رویکردهایی استفاده کرده‌اند که پیش‌تر نیز به کار برده شده‌اند. (مانند روش سالتون (۱۹۸۳)⁶) در پردازش اولیه، بردارهای کلمه-تعداد ساخته می‌شوند و ایست‌واژه‌ها از بردارها حذف می‌شوند. در زمان اجرا، یک کوئری (شامل پاراگراف یا متنی به زبان طبیعی) به سیستم داده می‌شود. این کوئری به یک نمایش برداری ترجمه می‌شود (این‌جا نیز ایست‌واژه‌ها حذف می‌شوند)، ریشه‌یابی می‌شوند و سپس با اسناد موجود در دیتابیس -به کمک یکی از روش‌های اندازه‌گیری شباهت مانند معیار کسینوسی، معیار جاکارد و معیار دایس- و روش‌های وزن‌دهی⁷ -شامل سه روش TF و دو نوع $TF*IDF$ ⁸ -مقایسه می‌شوند. پس از آزمایش سیستم و بررسی‌های بیشتر، نتیجه این بوده که معیار کسینوسی همراه با روش وزن‌دهی $TF*IDF$ نوع دوم بیشترین کارایی را داشته است.

الگوریتم ترازسازی HelioBlast یک الگوریتم ترازسازی نسبتاً جدید است که ورودی کاربر و هر سند را ترازسازی می‌کند. در این الگوریتم، از یک ماتریس جابجایی استفاده می‌شود که هر درج/حذف هزینه

⁴ Alignment algorithms

⁵ Text Similarity Searching (TSS)

بنا به گفته لوییس و همکاران (۲۰۰۶)، سیستم‌های TSS عموماً سندها را به صورت فهرستی از کلمات و فراوانی آن‌ها نمایش می‌دهند. این فهرست و تعداد کلمات بردارهای کلمه-تعداد یا بردار نامیده می‌شوند و می‌توانند با روش‌های مختلف با هم مقایسه شوند.

⁶ Salton G., *Introduction to Modern Information Retrieval*, 1983 New York McGraw-Hill

⁷ Weighting scheme

⁸ لوییس و همکاران (۲۰۰۶) استفاده از دو نوع $TF*IDF$ را گزارش کرده‌اند؛ نوع دوم $TF*IDF$ تنها تغییر کوچکی در ایده اصلی می‌دهد و مدعی می‌شود آستانه/مرزی وجود دارد که افزایش تکرار یک کلمه در سند نشانگر اهمیت بیشتر آن کلمه نیست. برای اعمال این ایده در روش وزن‌دهی، TF به یک تابع لگاریتمی (با پایه ۱/۶) داده می‌شود تا وزن کلماتی که فراوانی زیادی دارند به نوعی کنترل شود.

۱- دارد، امتیاز عدم تطبیق ۰ و امتیاز تطبیق برابر با وزن IDF آن کلمه است. دو الگوریتم ترازسازی وجود دارد: الف) ترازسازی کامل متنی: که در آن تمام سندهای دیتابیس (چکیده‌های MEDLINE) است با ورودی کاربر ترازسازی می‌شوند. ب) ترازسازی جمله‌ای: که جملات ورودی کاربر با هر جمله سند ترازسازی می‌کند، بیشترین مقادیر ترازسازی را ذخیره می‌کند و در نهایت، این مقادیر را با هم جمع می‌کند تا مقدار ترازسازی را برای کل سند محاسبه کند. روش دوم کارایی بهتری داشته است. دیتابیس این نرم‌افزار شامل حدود ۵۰ گیگابایت فایل اکس‌ام‌ال است که به ۱۳ گیگابایت متن قابل جستجو فشرده‌سازی شده و هر عنوان، چکیده و فهرست نویسندگان را با یک بردار کلمه-تعداد (ایست‌واژه‌ها حذف شده‌اند) نمایش داده شده است.

این سیستم توسط ده دانشجوی پزشکی تست شده است؛ از هر دانشجو خواسته شده یک پاراگراف/قطعه متن درباره موضوعی که درباره‌اش می‌دانند، بنویسند و به سیستم بدهند و ۳۰ سند اول نتایج را بررسی کنند. بر اساس نتایج این آزمایش، میانگین precision این سیستم $11/3 \pm 76/8$ گزارش شده است و میانگین بازیابی این سیستم $5/5 \pm 23/5$ گزارش شده است.

در زیر نمونه‌ای از ورودی و بخشی از خروجی این نرم‌افزار می‌بینید. همان‌طور که مشخص است هر سند شامل یک درصد شباهت است که این درصد شباهت در شناسایی سرقت علمی مورد استفاده قرار می‌گیرد.


نمونه‌ای از یک متن ورودی و بخشی از خروجی HelioBlast:

Ask HelioBLAST

Search in: ☒ Medline Would you like to add your own database? [Get in touch with us.](#)

Acquired immunodeficiency syndrome (AIDS) is a chronic, potentially life-threatening condition caused by the human immunodeficiency virus (HIV). By damaging your immune system, HIV interferes with your body's ability to fight the organisms that cause disease. HIV is a sexually transmitted infection (STI)

[Submit to HelioBLAST](#)



HelioBLAST

HelioBLAST results similar to your query

The 50 best matches found by HelioBLAST in 2.302 sec:

The placental barrier to maternal HIV infection. Score:0.482
by Anderson, V M
in Obstetrics and gynecology clinics of North America (1997)

Abstract: Infection with HIV destroys the immune system and causes acquired immunodeficiency syndrome (AIDS). Death results from common bacterial and opportunistic infections that are rare in persons with a healthy immune system. HIV infection frequently is a fatal sexually transmitted disease that can also be transmitted from an infected mother to her offspring. [More>>](#)

Medline PMID: [9430168](#)

HIV infection in women. Score:0.476
by Wenstrom, K D; Gall, S A
in Obstetrics and gynecology clinics of North America (1989)

Abstract: Acquired immunodeficiency syndrome (AIDS) cripples the immune system, making the patient susceptible to a variety of infections and disorders. AIDS is caused by direct blood contact with a transmissible agent, now known as the human immunodeficiency virus (HIV). The pathophysiology, clinical disease, and treatment of HIV infection in women are discussed. [More>>](#)

Medline PMID: [2687749](#)

Sexually transmitted diseases (STD) / reproductive tract infections (RTI) including acquired immunodeficiency syndrome (AIDS) / human immunodeficiency virus (HIV) infections among the women of reproductive age group: a review. Score:0.474
by Nahar, A; Azad, A K
in Journal of preventive and social medicine : JOPSOM : a bi-annual journal of the National Institute of Preventive and Social Medicine (1999)

No Abstract Available

Analysis Tools

A few tools to do more:

Find An Expert

Experts are potential reviewers, collaborators or competitors. Experts are identified from their publication history in this search.

1. [Trillo-Pazos, G](#) ★ ☆ ☆

Implicit Keywords

Implicit Keywords help identify concepts that were not originally mentioned in the query. Words are extracted from the 50 best matches found by HelioBLAST.

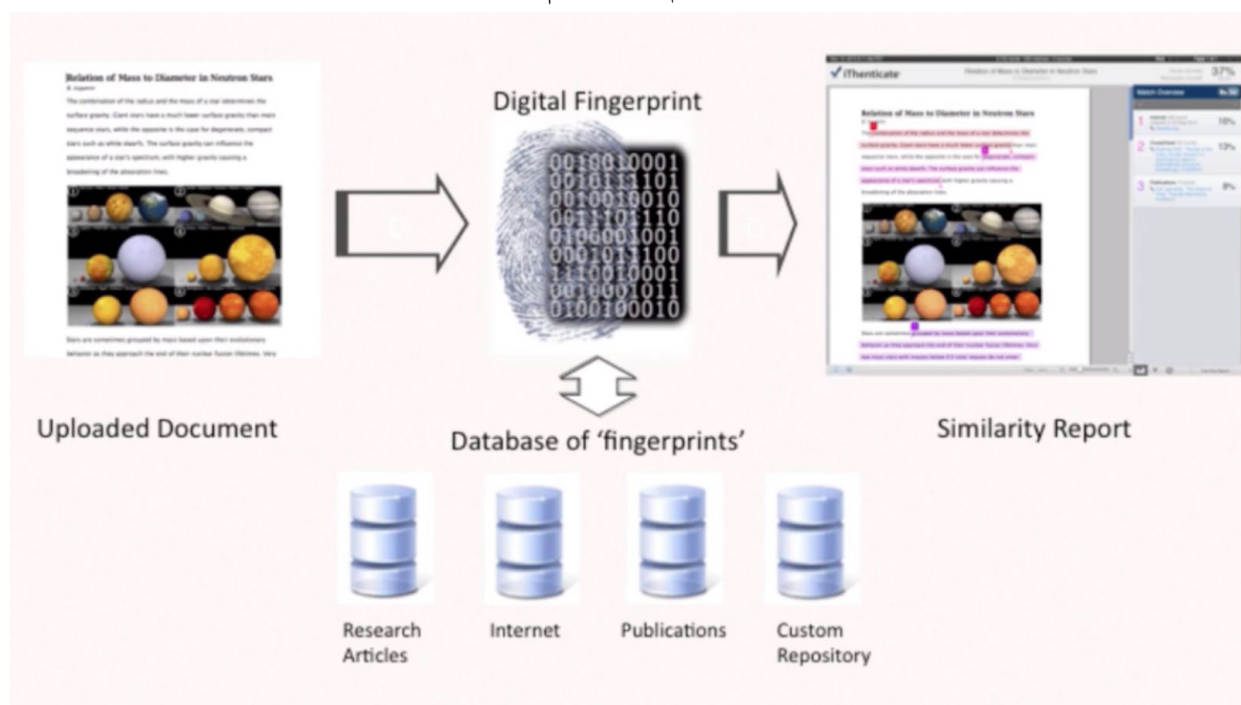
[View Implicit Keywords>>](#)

در ادامه به بررسی اجمالی نرم افزارهای دیگر شباهت یاب متن می پردازیم:

۲- **iThenticate**: این ابزار از معروفترین و پخته ترین ابزارهای بررسی سرقت علمی است و ژورنال های معتبری مثل اسپرینگر و الزویر از آن استفاده می کنند. ما نتوانستیم درباره جزئیات یا روش های دقیق مورد استفاده این نرم افزار چیزی پیدا کنیم (که با توجه به تجاری و معروف بودن آن منطقی به نظر می رسد). با این حال، در اسناد این نرم افزار اشارات کوچکی به روش های آن شده است.

iThenticate با الگوریتمی هر متن را به یک «اثر انگشت دیجیتال» تبدیل می کند؛ که با [متون] دیتابیس کاملی مورد مقایسه قرار می گیرد- درست مانند اثر انگشت انسان- و حتی کوچکترین آثار شباهت نیز از نظرش دور نمی ماند.⁹

در نمودار زیر، فرآیند تشخیص شباهت در این نرم افزار ترسیم شده است:



⁹. <https://www.ithenticate.com/hs-fs/hub/92785/file-227590694-pdf/docs/plagiarism-detection-misconceptions.pdf>

در زیر نمای کلی واسط کاربری این نرم افزار را می بینید:

Folders Settings Account Info Manage Users Welcome Jessica Gopalakrishnan | Logout Help

iThenticate®
Professional Plagiarism Prevention

Search Trash

My Folders
My Folders
My Documents
Kenneth Balib...
Trash

My Documents Documents Sharing Settings page 1 of 1

Title	Report	Author	Processed ↓	Actions
<input type="checkbox"/> Health Care and Evidence Base 1 part - 1,848 words	13%	William R. McMillan	Thu Jun 20, 2013 11:31am PDT	
<input type="checkbox"/> Medical and Health Care Reform 1 part - 1,818 words	12%	William R. McMillan	Thu Jun 20, 2013 11:24am PDT	
<input type="checkbox"/> A2088446-Ghost1.docx 1 part - 1,779 words	10%		Mon Mar 18, 2013 11:01am PDT	
<input type="checkbox"/> Diameter of Stars.docx 1 part - 179 words	95%		Wed Jan 30, 2013 12:14pm PST	

page 1 of 1

Submit a document
9,790 Documents remaining
[Upload a File](#)
[Zip File Upload](#)
[Multiple File Upload](#)
[Cut & Paste](#)
View: [Recent Uploads](#)

New folder
[New Folder](#)
[New Folder Group](#)

https://www.ithenticate.com/doc/report/10080811/kalibib...

در زیر، نمونه خروجی این نرم افزار در بررسی یک مقاله را می بینید. خطوط هایلايت شده خطوط مشکوک به سرقت علمی هستند و در سمت راست، منابع احتمالی این خطوط را می بینیم:

The Politics and Problems of Health Insurance

12 America's Health care System

The health care system was until the last few decades managed by fee for system i.e. people paid for services. Comparatively recently this has changed to one that is a managed care system although the brunt of it is still fee-for-service. Problems with the FFS are numerous including the fact that there is discrimination in health delivery with a great swaths of the population receiving inadequate or utter lack of care and with service being questionable and of limited value.

2 The 2010 Affordable Care Act will reform health insurance, over several years, meaning this law holds insurance companies less accountable, expands coverage for healthier adults, offers small-business tax credits, and provides access to insurance for uninsured Americans with pre-existing conditions.

11 Costs are held down by three kinds of services Health management organizations (HMOs), Independent practice Associations (IPA), and preferred Provider Organizations (PPO). Physicians are largely joining up with hospitals to provide services, and, in a manner that is different to most other nations where health insurance is either federally provided or funded by employer, medical insurance in the US is paid for by a combination of government, individuals, and employers.

Match Overview

Match	Source	Words	Percentage
1	Internet	43 words	2%
2	Internet	39 words	2%
3	Internet	32 words	2%
4	Internet	29 words	2%
5	Internet	28 words	1%
6	Internet	15 words	1%
7	Internet	14 words	1%
8	Internet	11 words	1%

در واقع، همان طور که انتظار می رود، این نرم افزار مثل یک موتور جستجوی بسیار پیشرفته و دقیق عمل می کند. دیتابیس آن نیز شامل تعداد زیادی مقالات ژورنالی، پایان نامه و صفحات وب است.¹⁰


۳- **Turnitin**: این ابزار یکی از ابزارهای معروف دیگر تشخیص سرقت در متن است و مشتریان آن بیشتر دانشگاه ها و موسسات آموزشی اند. دیتابیس این نرم افزار عمدتاً شامل صفحات وب، ژورنال ها و مقالات، و مقالات دانشجویی است.¹¹ در صفحه بعد تصویری از نمونه خروجی این نرم افزار بر یک متن ورودی دیده می شود:

¹⁰ <https://wiki.nus.edu.sg/display/cit/Turnitin+and+iThenticate+Comparison>

feedback studio Tessa Ruiz The Goliath of the Sea

The Goliath of the Sea

The majestic blue whale, the goliath of the sea, certainly stands alone within the animal kingdom for its adaptations beyond its massive size. At 30 metres (98 ft) in length and 190 tonnes (210 short tons) or more in weight, it is the largest existing animal and the heaviest that has ever existed. Despite their incomparable mass, aggressive hunting in the 1900s by whalers seeking whale oil drove them to the brink of extinction. But there are other reasons for why they are now so endangered.



Blue Whale - *Balaenoptera Musculus*

The blue whale's common name derives from bluish-hue that covers the upper side of its body, while its Latin designation is *Balaenoptera musculus*. The blue whale belongs to the Mysticeti suborder of cetaceans, also known as baleen whales, which means they have fringed plates of fingernail-like

Page: 1 of 2 Word Count: 517

Match Overview

49%

Rank	Source	Percentage
1	en.wikipedia.org	36%
2	animals.nationalgeogr...	7%
3	agaunews.com	6%

Text-only Report High Resolution

Want to learn more about Feedback Studio for your classroom? [Schedule a Consultation](#)

:Queuetext -۴

یکی از ابزارهای تشخیص سرقت علمی در متن است. دیتابیس این ابزار عمدتاً متشکل از منابع اینترنتی و کتاب‌های آنلاین است و شامل مقالات ژورنالی نمیشود. این ابزار، حتی اگر برخی لغات نیز تغییر کرده باشند، نیز تا حدی قادر به تشخیص سرقت علمی در متن است، با این حال false positive در تشخیص این نرم‌افزار نرخ بالایی دارد. این نرم‌افزار با هایلایت کردن بخش‌هایی از متن، عدم اصالت آن‌ها را مشخص می‌کند و با پلاگین مرجع‌دهی که دارد، کمک می‌کند تا کاربر منبع و مرجع بخش‌های هایلایت (فاقد اصالت) شده را مشخص کند.¹² در زیر دو نمونه از خروجی این نرم‌افزار را می‌بینید. تصویر اول نشانگر تمام بخش‌هایی است که با منابع دیگر مطابقت داشته‌اند و تصویر دوم، نشان می‌دهد که با انتخاب بخشی از متن، می‌توان جزئیات بیشتری درباره منبع منطبق با آن بخش پیدا کرد.

¹² <https://www.scribbr.com/plagiarism/best-plagiarism-checker/>

untitled

The atmosphere of Earth is composed of nitrogen (about 78%), oxygen (about 21%), argon (about 0.9%), carbon dioxide (0.04%) and other gases in trace amounts.[3] Oxygen is used by most organisms for respiration; nitrogen is fixed by bacteria and lightning to produce ammonia used in the construction of nucleotides and amino acids; and carbon dioxide is used by plants, algae and cyanobacteria for photosynthesis. The atmosphere helps to protect living organisms from genetic damage by solar ultraviolet radiation, solar wind and cosmic rays. The current composition of the Earth's atmosphere is the product of billions of years of biochemical modification of the paleoatmosphere by living organisms.

The 23S twisted circular form of ColE1 DNA has been isolated from Escherichia coli as a tightly associated DNA-protein complex with a sedimentation coefficient of approximately 24S. Treatment of this complex with pronase, trypsin, sodium dodecyl sulfate, Sarkosyl, or heat results in a conversion to a slower sedimenting form of 17S or 18S, as determined by centrifugation in neutral sucrose gradients. These treatments do not alter the sedimentation properties of noncomplexes supercoiled ColE1 DNA even in the presence of the ColE1-protein complex. Electron microscopic analyses indicate that the decrease in sedimentation rate of the ColE1-protein complex after treatment with these various agents is largely owing to an induced transition of ColE1 DNA from the supercoiled to the open circular state.

Counting rods and most abacuses have been used to represent numbers in a positional numeral system. With counting rods or abacus to perform arithmetic operations, the writing of the starting, intermediate and final values of a calculation could easily be done with a simple additive system in each position or column. This approach required no memorization of tables (as does positional notation) and could produce practical results quickly. For four centuries (from the 13th to the 16th) there was strong disagreement between those who believed in adopting the positional system in writing numbers and those who wanted to stay with the additive-

100% 13 matches from 3 sources

- en.wikipedia.org
<https://en.wikipedia.org/wiki/Atmosphere>
- pnas.org
<https://www.pnas.org/content/pnas/62/4/1159.full.pdf>
- en.wikipedia.org
https://en.wikipedia.org/wiki/Positional_notation

Help

100% 13 matches from 3 sources

Cite this source

100% similar en.wikipedia.org
https://en.wikipedia.org/wiki/Positional_notation

... Counting rods and most abacuses have been used to represent numbers in a positional numeral system. **With counting rods or abacus to perform arithmetic operations, the writing of the starting, intermediate and final values of a calculation could easily be done with a simple additive system in each position or column.** This approach required no memorization of tables (as does positional notation) and could produce pr ...

Help

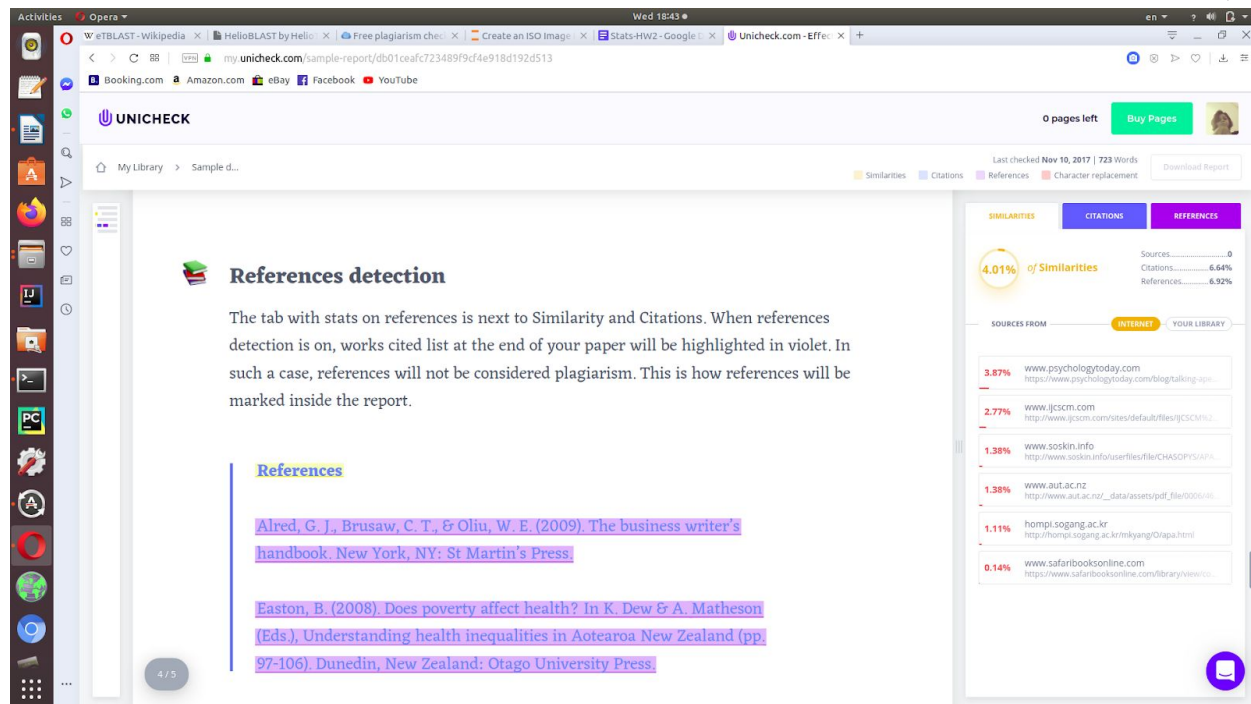
۵- Scribbr:

یکی از ابزارهای تشخیص سرقت در متن است. دیتابیس این ابزار بزرگ است و شامل متون وب، مقالات ژورنالی و نشریات است. این ابزار علاوه بر یافتن شباهت‌های دقیق متنی (تکرار عینی بخشی از متنی در متن دیگر)، تا حدی قادر به تشخیص متن بازنویسی شده (با کلمات اندکی متفاوت) نیز هست. گزارش آن نیز شامل بخش‌های هایلایت شده (متونی که احتمال دارد از جایی برداشته شده باشند) و فهرستی از منابع (متون شبیه به متن در حال بررسی) است.¹³ یک نمونه از خروجی این نرم‌افزار برای یک قطعه متن ورودی را در زیر می‌بینید:

The screenshot displays the Scribbr Plagiarism Checker interface. The main document, titled 'My-dissertation.docx', contains text with several green highlights indicating potential plagiarism. The highlighted text includes: '3.2 operationalization', '3.2.1 Dependent variable: Affiliation with the EU 253', 'Measuring affiliation with the EU', 'In the EVS dataset there is a question that asks the respondents to which of the geographical groups they would say they belong to first of all. The type of answer categories are: the', 'locality or town where you live, region of country where you live, country, Europe, the world as a whole. The questions above is recoded into one variable called "affiliationEU". This recoded variable shows if a respondent is more affiliated with the EU than with any other answer categories. A score of 1 shows that a respondent is most affiliated with the EU and a score of 0 shows that a respondent is more affiliated with either their locality or town, region', and 'locality or town where you live, region of country where you live, country, Europe, the world as a whole. The questions above is recoded into one variable called "affiliationEU". This recoded variable shows if a respondent is more affiliated with the EU than with any other answer categories. A score of 1 shows that a respondent is most affiliated with the EU and a score of 0 shows that a respondent is more affiliated with either their locality or town, region'. On the right side, the 'Sources Overview' panel shows an overall similarity of 28%. It lists a source from 'link.springer.com' with a 2% similarity. The source is identified as 'European Values Survey (EVS) Sense of belonging 1999-2002: To which of these geographical groups would you say you belong first of all? And next? 1995-1996/1990. To which of these geographical groups would you say you belong first of all? And the next? Answer categories: Locality or town where you live, state or region of country where you live, [country] as a whole, [continent], the world as a whole National pride: How proud are you to be [nationally]? 2010-2012/2005-2006/1999-2002/1995-1996. Answer categories - Very proud, quite proud, not very proud, not at all proud'. The panel also includes a 'View Full Text' link and an 'Exclude Source' button.

۶- Unicheck:

این نرم افزار نیز از دیگر نرم افزارهای شناخته شده برای تشخیص سرقت علمی و ادبی است. خروجی آن کم و بیش شبیه خروجی نرم افزارهای بالاست. نمونه خروجی این نرم افزار را در زیر می بینید:



۷- هم چنین، در جستجوی اولیه درباره سیستم های تجاری شباهت یاب متن، ای پی آی های بسیاری یافت می شود؛ برای مثال، Twinword، که یک برنامه تحلیل متنی است امکانات مختلفی از جمله تحلیل sentiment، تحلیل emotion، دسته بندی متن، شباهت یاب متنی و چندین امکان دیگر دارد. شباهت یاب متنی Twinword، میزان شباهت دو واژه، جمله یا پاراگراف را اندازه می گیرد. شکل ۱ تصویری از خروجی نسخه رایگان این برنامه است.

Rosette Text Analytics برنامه و ابزار دیگری برای تحلیل ویژگی های متنی است که از word embedding استفاده می کند و یکی از امکانات آن یافتن عبارات مشابه عبارت ورودی، در زبان های مختلف است. برای مثال، در صورتی که عبارت ورودی واژه «spy» و زبان های مطلوب ما شامل زبان ژاپنی، آلمانی و اسپانیایی باشد، بخشی از خروجی این سیستم (مشابه ترین ده عبارت به spy در زبان ژاپنی) را در شکل ۱ خواهد بود:

```

{
  "similarTerms": {
    "jpn": [
      {
        "term": "スパイ",
        "similarity": 0.5544399
      },
      {
        "term": "諜報",
        "similarity": 0.46903181
      },
      {
        "term": "MI6",
        "similarity": 0.46344957
      },
      {
        "term": "殺し屋",
        "similarity": 0.41098994
      },
      {
        "term": "正体",
        "similarity": 0.40109193
      },
      {
        "term": "ブレデター",
        "similarity": 0.39433435
      },
      {
        "term": "レンズマン",
        "similarity": 0.3918637
      },
      {
        "term": "S.H.I.E.L.D.",
        "similarity": 0.38338536
      },
      {
        "term": "サーシャ",
        "similarity": 0.37628397
      },
      {
        "term": "黒幕",
        "similarity": 0.37256041
      }
    ]
  }
}

```

Fig. 1 - Rosette Text Analysis

twinword.com/api/text-similarity.php

Text Similarity

Evaluate the similarity of two words, sentences, or paragraphs.

Free Tool & API Demo

First, input some text:

Genzeb Shumi Regasa is an Ethiopian-born middle distance runner who competes internationally for Bahrain. She won the 1500 metres gold medal at the Asian Athletics Championships in 2011 and the Asian Indoor Athletics Championships in 2012.

Second, input some other text to compare with:

Elena Romagnolo is an Italian steeplechaser, middle and long-distance runner. She is the national record holder in the 3000 metres steeplechase, but now competes mainly in the 5000 metres.

Evaluate

Results:

```

{
  "similarity": 0.40711291452576,
  "value": 733252.67699948,
  "version": "4.0.0",
  "author": "twinword inc.",
  "email": "help@twinword.com",
  "result_code": "200",
  "result_msg": "Success"
}

```

Fig. 2 - Twinword API

DigitalOwl نرم افزار دیگری است چند سرویس متنی (دسته بندی، شباهت یابی و تحلیل گروه نحوی) دارد. در شکل ۳ خروجی نمونه سرویس شباهت یابی آن را می بینید:

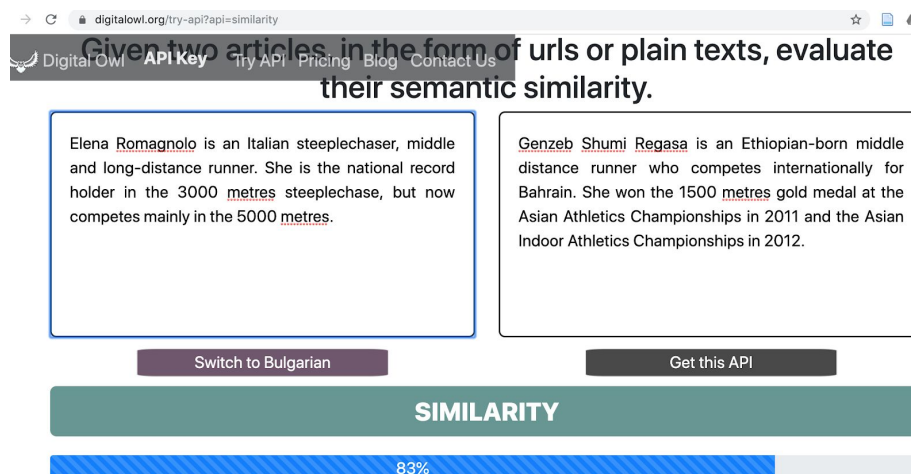


Fig. 3 - Digital Owl Text Similarity

سوال ۲-

گام ۱: ماتریس کلمه سند $X_{6 \times 2}$

(کلمات «شد»، «شده» و «نمایید»، بر اساس فهرست ایست‌واژه‌های این تمرین، از سندها حذف شده‌اند)

	d1	d2	q
دیروز	1	0	0
خبری	1	1	1
منتشر	1	1	0
به	0	1	0
بخش	0	1	0
توجه	0	1	1

گام ۲ (به کمک numpy): تجزیه $X = U \sum V^T$ ب)

```
>>> u
array([[ -0.21052722,  0.64048711, -0.36294535,  0.37136021,  0.37136021,
         0.37136021],
       [ -0.55116742,  0.24464431, -0.43435856, -0.3863057 , -0.3863057 ,
        -0.3863057 ],
       [ -0.55116742,  0.24464431,  0.79730392,  0.01494549,  0.01494549,
         0.01494549],
       [ -0.3406402 , -0.39584281, -0.12098178,  0.7904534 , -0.2095466 ,
        -0.2095466 ],
       [ -0.3406402 , -0.39584281, -0.12098178, -0.2095466 ,  0.7904534 ,
        -0.2095466 ],
       [ -0.3406402 , -0.39584281, -0.12098178, -0.2095466 , -0.2095466 ,
         0.7904534 ]])

>>>
>>> s
array([2.49721204, 1.32813103])

>>> vt
array([[ -0.52573111, -0.85065081],
       [ 0.85065081, -0.52573111]])

>>>
```

u همان U است. s همان Σ است و vt همان V^T است.
گام ۳: کاهش بعد

k=2 می گیریم (زیرا ۳ مقدار ویژه ی دیگر برابر با صفرند) پس:

$$U_{6 \times 2} =$$

$$\begin{bmatrix} -0.211 & 0.640 \end{bmatrix}$$

$$\begin{bmatrix} -0.551 & 0.245 \end{bmatrix}$$

$$\begin{bmatrix} -0.551 & 0.245 \end{bmatrix}$$

$$\begin{bmatrix} -0.341 & -0.396 \end{bmatrix}$$

$$\begin{bmatrix} -0.341 & -0.396 \end{bmatrix}$$

$$\begin{bmatrix} -0.341 & -0.396 \end{bmatrix}$$

$$\Sigma_{2 \times 2} =$$

$$\begin{bmatrix} 2.497 & 0. \end{bmatrix}$$

$$\begin{bmatrix} 0. & 1.328 \end{bmatrix}$$

$$V^T_{2 \times 2} =$$

$$\begin{bmatrix} -0.526 & -0.851 \end{bmatrix}$$

$$\begin{bmatrix} 0.851 & -0.526 \end{bmatrix}$$

گام ۴: محاسبه بردار سند یا بردار کلمه

بردار هر کلمه: ردیف های ماتریس $\Sigma_{2 \times 2}$ $U_{6 \times 2}$

بردار هر سند: ستون های ماتریس $\Sigma_{2 \times 2} V_{2 \times 2}^T$

که به ترتیب در زیر آمده است:

$$U_{6 \times 2} \Sigma_{2 \times 2} =$$

$$\begin{bmatrix} [-0.526 & 0.851] \\ [-1.376 & 0.325] \\ [-1.376 & 0.325] \\ [-0.851 & -0.526] \\ [-0.851 & -0.526] \\ [-0.851 & -0.526] \end{bmatrix}$$

$$\Sigma_{2 \times 2} V_{2 \times 2}^T =$$

$$\begin{bmatrix} [-1.313 & -2.124] \\ [1.13 & -0.698] \end{bmatrix}$$

که در واقع به معنی زیر است:

-0.526	دیروز
0.851	

-1.376	خبری
0.325	

-1.376	منتشر
0.325	

-0.851	به
-0.526	

-0.851	بخش
-0.526	

-0.851	توجه
-0.526	

-1.313	d1
1.13	

-2.124	d2
-0.698	

۵- گام (استفاده): محاسبه شباهت پرسش «توجه خبری» به اسناد:

محاسبه بردار پرسش →

(۱) می‌توانیم میانگین بگیریم:

$$q1 = ([-0.851, -0.526] + [-1.376, 0.32]) / 2 = [-1.1135, -0.1004]$$

(۲) هم‌چنین می‌توانیم از فرمول $q = \sum^{-1} U^T d_{new}$ استفاده کنیم که در این صورت:

$$\sum^{-1} = \begin{bmatrix} 0.4 & 0. \\ 0. & 0.753 \end{bmatrix}$$

$$U^T =$$

$$\begin{bmatrix} -0.211 & -0.551 & -0.551 & -0.341 & -0.341 & -0.341 \end{bmatrix}$$

$$\begin{bmatrix} 0.64 & 0.245 & 0.245 & -0.396 & -0.396 & -0.396 \end{bmatrix}$$

$$d_{new} = [0, 1, 0, 0, 0, 1]$$

پس:

$$q_2 = [-0.357 \quad -0.114]$$

محاسبه شباهت کسینوسی با سندهای ۱ و ۲ →

$$\text{cosineSim}(d_1, q_1) = d_1 \cdot q_1 / |d_1| |q_1| = 0.696$$

$$\text{cosineSim}(d_2, q_1) = d_2 \cdot q_1 / |d_2| |q_1| = 0.974$$

یا

$$\text{cosineSim}(d_1, q_2) = d_1 \cdot q_2 / |d_1| |q_2| = 0.524$$

$$\text{cosineSim}(d_2, q_2) = d_2 \cdot q_2 / |d_2| |q_2| = 0.9999596996393096$$

سوال ۳-

کلاسی به نام SimilarityDetector ساختیم که آدرس پیکره را برای ساختن شی و مقداردهی اولیه دریافت می کند. این کلاس دو ویژگی دارد: IDF کلمات پیکره (idf) و آدرس فایل پیکره. در زیر توضیح مختصری درباره هر متد می دهیم:

- متد `_normalize`: این متد نرمال سازی متن را انجام می دهد. پارامتر `text` متن ورودی است، پارامترهای `removePunc` و `removeStopWords` که مقادیر بولی می پذیرند، به ترتیب، علائم سجاوندی و ایست واژه ها را از متن حذف می کنند.
- متد `removeStopwords`: این متد در متد `_normalize` فراخوانی می شود و وظیفه حذف ایست واژه ها را بر عهده دارد.
- متد `_prepareSentenceSet`: پیکره را از فایل می خواند و در اطلاعات آن را در لیستی از تاپل ها برمی گرداند؛ به شکل زیر:
$$[(s11, s21, similarity), (s12, s22, similarity), \dots]$$
- متد `outputNormalizeCorpus`: اسم فایل خروجی را دریافت می کند و جملات پیکره را، بعد از نرمال سازی و حذف ایست واژه ها، در فایل خروجی می ریزد.
- متد `selectLexicon`: واژگان پیکره را (بعد از نرمال سازی پیکره و حذف ایست واژه ها) در فایل `Dic.txt` می ریزد.
- متد `computeTFSimilarity`: دو پارامتر ورودی `s1` و `s2` معادل جمله اول و جمله دوم هستند. این تابع ابتدا دو جمله ورودی را نرمال می کند، تقطیع می کند و شباهت کسینوسی و شباهت جاکاردی آن ها را بر اساس بردار `TF` نرمال شده آن ها محاسبه می کند و شباهت ها را به صورت یک تاپل دوتایی برمی گرداند.
- متد `computeIDF`: پیکره نرمال شده را به عنوان ورودی می گیرد و `IDF` هر کلمه را محاسبه می کند و در ویژگی `idf` آبجکت نگهداری می کند.
- متد `computeTFIDFSimilarity`: مانند متد `computeTFSimilarity`. دو جمله ورودی می گیرد و شباهت را بر اساس معیار کسینوسی و جاکارد، به کمک روش وزن دهی `TFIDF` محاسبه می کند.

- متدهای `computeAverageTFSimilarity` و `computeAverageTFIDFSimilarity`: این دو متد، به ترتیب با روش‌های وزن‌دهی TF و TFIDF و معیار شباهت جاکارد و کسینوسی، میانگین شباهت را در پیکره اندازه می‌گیرند.
- متد `computeCorrelationCoefficient`: این متد به کمک کتابخانه `numpy` ضریب همبستگی نتایج را با داده واقعی محاسبه می‌کند.
- متد `computeCorrelationCoefficientFromScratch`: این متد نیز کار متد قبل را انجام می‌دهد؛ منتها بدون کمک کتابخانه `numpy` ضریب همبستگی را محاسبه می‌کند.

(الف)

پیکره نرمال‌شده (یکسان‌سازی کاراکترها و حذف ایست‌واژه‌ها و کاراکترهای نالازم) در فایل `normalizedSimilarityCorpusSample.csv` ذخیره شدند. فایل `Dic.txt` نیز حاوی واژگان پیکره است.

(ب)

ر.ک. متد `computeTFSimilarity` و `computeAverageTFSimilarity`.

میانگین شباهت کسینوسی با TF	میانگین شباهت جاکارد با TF
۰/۶۱۲	۰/۱۹۶

*** شباهت جاکارد با استفاده از فرمول جاکارد موجود در اسلایدها یعنی

$$\text{jaccard similarity}(d1, d2) = \frac{d1.d2}{|d1| + |d2| - d1.d2}$$

محاسبه شده است. با این حال، این معیار شباهت را می‌توان به شیوه دیگری نیز محاسبه کرد؛ برای مثال پکیج `scipy` شباهت جاکارد را با فرمول زیر محاسبه می‌کند:

$$\text{jaccard similarity}(d1, d2) = \sum (a_i == b_i \neq 0) / \sum (a_i \neq 0 \text{ or } b_i \neq 0)$$

که البته بیشتر مناسب بردارهای باینری است؛ زیرا در این روش، اگر a_i مثلا برابر با ۱ باشد، b_i چه ۱/۵ باشد و چه ۱۰۰/۵ فرقی ندارد و اشتراک در هر دو مورد صفر است.

(ج)

ر.ک. `computeTFIDFSimilarity` و `computeAverageTFIDFSimilarity`.

میانگین شباهت TFIDF	میانگین شباهت کسینوسی با TFIDF
۰/۳۶۲	۰/۴۷

(د)

ر.ک. متد `computeCorrelationCoefficientFromScratch` و/یا

`computeCorrelationCoefficient`.

	TF	TFIDF
کسینوسی	0.645	0.768
جاکارد	0.589	0.757

ضریب همبستگی

شباهت کسینوسی با روش وزن‌دهی TFIDF بیشترین همبستگی را با داده اصلی دارند و در رتبه دوم، شباهت جاکارد با روش وزن‌دهی TFIDF است. در نتیجه، روش وزن‌دهی TFIDF برتری قابل توجهی

نسبت به روش وزن‌دهی TF دارد؛ این نتیجه منطقی به نظر می‌رسد؛ زیرا در TFIDF، ما دو ویژگی از متن استخراج کرده‌ایم و اطلاعات بیشتری را در محاسبه شباهت دخیل کرده‌ایم در حالی که در TF، تنها یک ویژگی از متن در دست داریم.

در روش وزن‌دهی TF نیز مانند TFIDF، ضریب همبستگی شباهت کسینوسی در مقام اول و شباهت جاکارد در مقام دوم است. در نتیجه، می‌توان گفت بین دو معیار محاسبه شباهت نیز، معیار کسینوسی عملکرد بهتری داشته است.

سوال ۴-

کلاس NaiveBayesTextClassifier شامل متدهای:

- (۱) `normalize_`: برای نرمال سازی (مانند سوال ۳)
 - (۲) `removeStopwords_`: برای حذف ایست واژه ها (مانند سوال ۳)
 - (۳) `splitTrainTest`: برای جداسازی دادگان آموزش و آزمون. دادگان آموزش در پوشه ZebRa (که در آدرس روت پروژه قرار دارد) باقی می ماند و دادگان تست به پوشه ای به نام Test منتقل می شوند.
 - (۴) `mergeCategoryFiles`: برای مرج کردن و به هم چسبانیدن فایل های یک موضوع. هر موضوع هشت فایل برای آموزش سیستم دارد؛ این هشت فایل یکی می شوند و در پوشه MergedCats قرار می گیرند.
 - (۵) `calculateConditionalProbabilities`: این متد احتمال های شرطی مورد نیاز برای مدل بیزی ساده را محاسبه می کند.
 - (۶) `dumpConditionalProbsToJson`: این متد احتمال های محاسبه شده در متد شماره ۵ را به یک فایل جیسونی منتقل می کند. کلیدهای این فایل موضوع ها (۷ کلید) و مقادیر این کلیدها یک دیکشنری است؛ این دیکشنری حاوی کلمات و احتمال های آن ها به شرط موضوعات است. این احتمال ها در فایل `probs.json` ذخیره شده اند.
 - (۷) `loadConditionalProbsToJson`: این متد احتمال ها را از فایل `probs.json` می خواند.
 - (۸) `classifyTestDoc`: این متد با گرفتن آدرس فایل تست و مدل احتمالاتی شرطی، موضوع سند را مشخص می کند.
 - (۹) `evaluate`: این متد ماتریس درهم ریختگی را می سازد و چاپ می کند. صحت (precision) و بازخوانی کلی را با دو شیوه میانگین گیری ریز و میانگین گیری درشت محاسبه می کند و هم چنین، صحت و بازخوانی را برای هر کلاس به صورت مجزا گزارش می دهد.
- ** هنگام ساختن مدل و هم چنین هنگام دسته بندی سند تست، کلمات به کمک ریشه یاب کتابخانه هضم ریشه یابی می شوند و بعد احتمال ریشه کلمات محاسبه می شود.**

(الف)

با استفاده از متدهای بالا واژگان استخراج شد و احتمال‌ها محاسبه شد. در صورتی که کلمه‌ای در داده تست باشد که در واژگان موجود نباشد، آن کلمه کنار گذاشته می‌شود.

(ب)

خروجی متد evaluate در زیر آمده است؛ این متد وظیفه ساختن ماتریس درهم‌ریختگی و محاسبه صحت و بازخوانی -به صورت کلی با میانگین‌گیری ریز و درشت و همچنین به تفکیک کلاس- را به عهده دارد. اندیس‌های i و j ماتریس نیز که در یک دیکشنری ذخیره شده‌اند، چاپ شده‌اند. یعنی خانه اول سطر و ستون ماتریس ما کلاس «ادیان»، خانه دوم کلاس «مسائل راهبردی»، خانه سوم کلاس «ورزشی» و ... است. صحت و بازخوانی کلی با دو روش میانگین‌گیری ریز و درشت نیز در انتها آمده است.

Matrix Indices:

{'ادیان': 0, 'مسائل راهبردی ایران': 1, 'ورزشی': 2, 'سیاسی': 3, 'اقتصادی': 4, 'فناوری': 5, 'اجتماعی': 6}

Confusion Matrix:

```
[[2. 0. 0. 0. 0. 0. 0.]
 [0. 1. 0. 0. 0. 1. 0.]
 [0. 0. 2. 0. 0. 0. 0.]
 [0. 0. 0. 2. 0. 0. 0.]
 [0. 0. 0. 1. 1. 0. 0.]
 [0. 0. 0. 0. 0. 2. 0.]
 [0. 0. 0. 0. 0. 0. 2.]]
```

ادیان:

Precision: 1.000 Recall: 1.000 F-Score: 1.000

مسائل راهبردی ایران:

Precision: 1.000 Recall: 0.500 F-Score: 0.667

ورزشی:

Precision: 1.000 Recall: 1.000 F-Score: 1.000

سیاسی:

Precision: 0.667 Recall: 1.000 F-Score: 0.800

اقتصادی:

Precision: 1.000 Recall: 0.500 F-Score: 0.667

فناوری:

Precision: 0.667 Recall: 1.000 F-Score: 0.800

اجتماعی:

Precision: 1.000 Recall: 1.000 F-Score: 1.000

Overall Micro Precision: 0.857

Overall Micro Recall: 0.857

Overall Micro F-score: 0.857

Overall Macro Precision: 0.905

Overall Macro Recall: 0.857

Overall Macro F-score: 0.880