 **بازیابی پیشرفته اطلاعات**

مدرس: دکتر بیگی

شماره گروه: 5

دانشکده مهندسی کامپیوتر تهیه‌کنندگان: نیما جمالی – سپهر فعلی – سینا کاظمی

گزارش فاز دوم پروژه

فهرست مطالب

[پیش‌پردازش اولیه 3](#_Toc56193670)

[پیش‌پردازش مستندات انگلیسی 3](#_Toc56193671)

[کلمات پرتکرار مستندات انگلیسی 4](#_Toc56193672)

[پیش‌پردازش مستندات فارسی 5](#_Toc56193673)

[کلمات پرتکرار مستندات فارسی 5](#_Toc56193674)

[نمایه‌سازی 7](#_Toc56193675)

[نمایه bigram 7](#_Toc56193676)

[نمایه positional 7](#_Toc56193677)

[توابع درج و حذف مستندات 8](#_Toc56193678)

[توابع ذخیره‌سازی و لود نمایه‌ها 10](#_Toc56193679)

[فشرده‌سازی نمایه‌ها 11](#_Toc56193680)

[فشرده‌سازی با variable byte 11](#_Toc56193681)

[فشرده‌سازی با gamma code 12](#_Toc56193682)

[ذخیره‌سازی و لود نمایه‌ها 13](#_Toc56193683)

[اصلاح پرسمان 14](#_Toc56193684)

[استفاده از نمایه bigram و معیار جاکارد برای یافتن کلمات مشابه 14](#_Toc56193685)

[تابع محاسبه‌ی فاصله ویرایش دو کلمه 15](#_Toc56193686)

[نمایش پرسمان اصلاح‌شده 15](#_Toc56193687)

[جستجو و بازیابی اسناد 16](#_Toc56193688)

[جستجوی ltc-lnc پرسمان 16](#_Toc56193689)

[جستجوی proximity با اندازه‌ی پنجره‌ی داده شده در ورودی 17](#_Toc56193690)

[نحوه‌ی تقسیم وظایف 19](#_Toc56193691)

[نیما جمالی 19](#_Toc56193692)

[سپهر فعلی 19](#_Toc56193693)

[سینا کاظمی 19](#_Toc56193694)

[مراجع 20](#_Toc56193695)

# ساختن فضای برداری برای دسته‌بندها

## پیش پردازش

## پیاده‌سازی فضای برداری tf-idf به روش ntn

# پیاده‌سازی دسته‌بندها

## Naïve Bayes

## k-NN

## SVM

## Random Forest

# بهبود سیستم بازیابی فاز اول پروژه

## تابع اجماع نظر دسته‌بندها برای تعیین برچسب هر مستند

## جستجو بر اساس دسته

# ارزیابی نهایی

## پیاده‌سازی تابع find\_metric

## معیارهای ارزیابی برای Naïve Bayes

## معیارهای ارزیابی برای k-NN

## معیارهای ارزیابی برای SVM

## معیارهای ارزیابی برای Random Forest

# نحوه‌ی تقسیم وظایف

وظایف اختصاص یافته به هر فرد به شرح زیر بود:

## نیما جمالی

1-

## سپهر فعلی

1-

## سینا کاظمی

1-

# مراجع