چالش های تحلیل متن فارسی:

>> پیش پردازش صحیح. نرمالایز کردن

>> اینکه در چه سطحی توکن کنیم خیلی مهمه. توکن ها نه زیاد باید طویل باشن نه خیلی به سوی اتومیک پیش رفتن. مثلا به این جمله توجه کنید:

دکتر حسن روحانی به نیویورک سفر می کند.

در این جمله، اگر توکنایز بتواند "دکتر حسن روحانی " را دتکت کند برای مدل سازی بسیار خوب است. ولی اگر هر کدام ازکلمات را یک توکن در نظر بگیرد کمک چندانی نمیکند.

>> چالش دیگر چالش تشخیص افعال ترکیبی است. مثلا : سفر می کند. غذا می خورد.

راه حل: روکیرد فازی

چالش های tf-idf

چالش 1:

بهتره قبل از دادن دیتا به این الگوریتم، با روش خودمون توکنایز کنیم. جمله زیر رو در نظر بگیرید:

دکتر روحانی به نیویورک سفر می کند.

اگر توکن کردن ما از قبل به این شکل بوده باشه:

"دکتر روحانی"، "به نیویورک"، "سفر می کند"

خیلی بهتر میتونه عبارات رو اسکور دهی کنه. مثلا میگرده دنبال تکرار ترم "دکتر روحانی"

راه حل:

من ابتدا کل ستون کلمات رو برداشتم. خوبی این دیتا ست اینه که توکنایز کردنش در سطح عبارت. مثلا همون دکتر روحانی. بعدش به عنوان لیست خوندمش. لیستی از کلمات. خوب حالا داکیومنت و جملات رو از توش استخراج کردم. هر جمله یک سند. چطوری؟ توی اون ستون گشتم دنبال punctutation ها. گفتم هرجا از این پانک ها دیدی ینی جمله تموم شده.

چالش 2:

تو این مساله خاص که ما یه دیتای لیبل شده داریم و میخوایم ازش ویژگی استخراج کنیم، آیا بعد از محاسله tf\_idf باید امتیاز هر ترم رو بزاریم تو همون دیتا ست اصلی برای طبقه بندی؟ دیتای تست تکلیفش چی میشه؟

چالش 3: ایست واژه ها باید تو این فرآیند باشند یا خیر؟