Luppar News-Rec: Um Recomendador Inteligente de notícias

Antonio Alex de Souza
Orientador: Prof. Dr. José Everardo Bessa Maia

Mestrado Acadêmico em Ciência da Computação
Universidade Estadual do Ceará (UECE)

20 de agosto de 2019

Roteiro

INTRODUÇÃO

Sistemas de Recomendação (SR) geram recomendações individualizadas.

Difere do cenário geral de recomendação em:
- Ciclo de Vida, Interesse, Eventos, Dados.

Um Sistema de Recomendação de Notícias (SRN) pode usar uma ou mais das abordagens:

- Baseada em Conteúdo*;
- Baseada em Contexto;
- Baseada em Filtragem Colaborativa;
- Híbrida.

OBJETIVOS:

- **Objetivo Geral**
  Construir e avaliar um SRN competitivo com o estado da arte, utilizando representações *embedding* de documentos.

- **Objetivos Específicos**
  - Especificar o contexto da aplicação e desenvolver uma estrutura para o SRN;
  - Obter uma proposta de representação *embedding* de documentos competitiva e comparar seu desempenho com representações estado da arte;
  - Construir um subsistema classificador monorrótulo e multirrótulo de notícias e avaliar combinações de algoritmos de classificação e representações de documentos para selecionar a melhor proposta para o SRN;
  - Implementar o SRN com interface de usuário para teste e demonstração online, incluindo aquisição, pré-processamento, armazenamento, classificação e distribuição de notícias.

Roteiro

BAG-OF-WORDS (BOW) E TF-IDF

1. *I love dogs.*
2. *I hate dogs and knitting.*
3. *Knitting is my hobby and my passion.*

	i	love	dogs	hate	and	knitting	is	my	hobby	passion
Doc 1	1	1	1							
Doc 2	1		1	1	1	1				
Doc 3					1	1	1	2	1	1

$$\text{tf-idf}_{t,d} = (1 + log\ \text{tf}_{t,d}).log\frac{N}{\text{df}_t}$$

	i	love	dogs	hate	and	knitting	is	my	hobby	passion
Doc 1	0.18	0.48	0.18							
Doc 2	0.18		0.18	0.48	0.18	0.18				
Doc 3					0.18	0.18	0.48	0.95	0.48	0.48

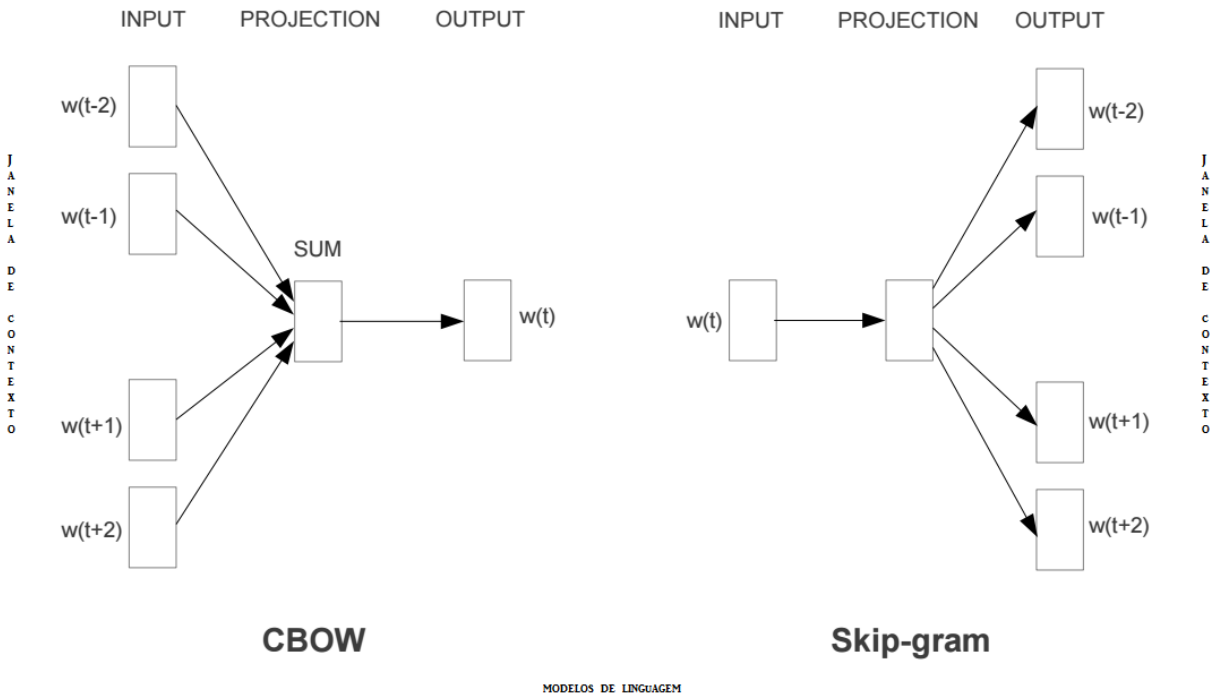
Problemas: alta dimensionalidade, ambiguidade semântica.

WORD EMBEDDINGS

Representar um termo por um vetor de números reais, denso e de tamanho arbitrário.

Word2Vec - Sentido semântico dos termos (próximos no vetor);

FastText - Informação morfológica dos termos.



WORD EMBEDDINGS

A partir da representação *embedding* dos termos é necessário obter uma **representação de documentos**. Normalmente um documento (notícia) é representado pelo vetor média dos vetores dos termos que o compõe (*Word2Vec*).[1]

	-	-	-	-	-	-	-	-	...	-
Doc 1	-0.82	-0.04	0.67	0.35	-0.30	-0.66	1.25	-1.11	...	1.63
Doc 2	1.32	0.98	-0.10	-1.48	0.23	-0.71	-0.87	-0.18	...	-0.89
Doc 3	-0.51	1.54	-0.56	0.67	-0.32	-0.42	1.99	0.91	...	-1.21

[1] KIM, H. K.; KIM, H.; CHO, S. Bag-of-concepts: Comprehending doc. rep. through clustering words in distr. repres.Neurocomputing, Elsevier, 2017.

ABORDAGEM PROPOSTA (E2V-IDF) [2]

Representa um documento pela média dos vetores dos seus termos, ponderando cada vetor de termo pelo IDF (Inverso da Frequência nos Documentos) do termo na coleção.

A ponderação *IDF* é dada pela fórmula:

$$IDF(w) = log\left(\frac{N}{df_t}\right) + 1 \quad (1)$$

onde w é o termo (palavra), N é o numero total de documentos da coleção e df_t é o número de documentos em que o termo w ocorre.

[2] Alex Souza e Everardo B. Maia. Agente Inteligente para Classificação de Notícias por Assunto. Anais do Computer on the Beach (2019)

ABORDAGEM PROPOSTA (E2V-IDF)

Algoritmo 1: Pseudocódigo da Representação de Documento E2V-IDF

Data: DOCUMENTOS (Notícias), VETOR *EMBEDDING* (Vetor de cada termo da coleção)

Result: Vetor E2V-IDF de cada Documento

while *Existir Termos na Coleção* **do**

  Calcula_IDF_Termo() //Monta o vetor: wIDF (Termo e IDF do Termo);

end

while *Existir Documentos na Coleção* **do**

  while *Existir Termos no Documento* **do**

    E2V-IDF(documento) = Média (VETOR *EMBEDDING*(termo) * wIDF(Termo));

  end

end

ROTEIRO

CLASSIFICADORES

- ***KNN (K-Nearest Neighbor)***;
  - k = 5.
  - Distância Euclidiana
- ***SVM (Support Vector Machines)***;
- ***DTree (Decision Tree)***;
- ***RF (Random Forest)***.

CLASSIFICADORES - MÉTRICAS

Precision - fração dos resultados retornados que são relevantes.

$$P = \frac{VP}{VP + FP} \tag{2}$$

Recall - quantos resultados realmente relevantes são retornados.

$$R = \frac{VP}{VP + FN} \tag{3}$$

F1-Score (F1)* - média ponderada do *precision e recall*

$$F1 = \frac{2 * P * R}{P + R} \tag{4}$$

Variância, Curva ROC, Curva UAC e RMSE também serão utilizadas para análises.

ROTEIRO

SELEÇÃO - REPRESENTAÇÃO *versus* CLASSIFICADOR

As avaliações ocorreram em sequência combinando:
- **6 coleções de documentos**: *Reuters-21578 (R8)*, *Reuters-21578 (R52)*, *Reuters RCV1*, *20 Newsgroups*, Z5News e Z5NewsBrasil;
- **4 classificadores**: SVM, KNN, DT e RF;
- **6 representações de documentos**: BoW, TF-IDF, *Word2Vec*, *Word2Vec* (`E2V-IDF`), *FastText* e *FastText* (`E2V-IDF`).

SELEÇÃO - REPRESENTAÇÃO *versus* CLASSIFICADOR

Coleções de Referência (Representações: BoW e TF-IDF)				
Modelo	**R8**	**R52**	**RCV1**	**20 News**
SVM(RBF)+BoW	0,9080 (0,0003)	**0,8675 (0,0003)**	0,6845 (0,0005)	0,5904 (0,0138)
SVM(RBF)+TFIDF	0,8870 (0,0004)	0,8573 (0,0004)	0,5451 (0,0002)	**0,6674 (0,0161)**
KNN+BoW	0,8965 (0,0003)	0,8388 (0,0003)	0,6989 (0,0002)	0,2889 (0,0044)
KNN+TFIDF	0,8795 (0,0005)	0,8509 (0,0004)	**0,7672 (0,0002)**	0,6397 (0,0089)
DT+BoW	0,9101 (0,0003)	**0,8615 (0,0002)**	**0,7180 (0,0003)**	0,5913 (0,0068)
DT+TFIDF	**0,9105 (0,0002)**	0,8531 (0,0004)	**0,7220 (0,0001)**	0,5760 (0,0082)
RF+BoW	**0,9208 (0,0003)**	0,8504 (0,0004)	0,6600 (0,0003)	**0,6716 (0,0112)**
RF+TFIDF	**0,9294 (0,0003)**	**0,8592 (0,0003)**	0,6789 (0,0003)	**0,6621 (0,0107)**

Tabela 6 – Resultados - F1-Score - Média (Variância)

SELEÇÃO - REPRESENTAÇÃO *versus* CLASSIFICADOR

Coleções próprias (Representações: BoW e TF-IDF)		
Modelo	**Z5 News**	**Z5 News Brasil**
SVM(RBF)+BoW	**0,7873 (0,0008)**	**0,8025 (0,0007)**
SVM(RBF)+TFIDF	0,7707 (0,0007)	**0,7744 (0,0006)**
KNN+BoW	0,6957 (0,0001)	0,6648 (0,0016)
KNN+TFIDF	0,7469 (0,0001)	0,7469 (0,0016)
DT+BoW	0,7318 (0,0007)	0,6825 (0,0037)
DT+TFIDF	0,7232 (0,0005)	0,6667 (0,0039)
RF+BoW	**0,7818 (0,0004)**	**0,7651 (0,0020)**
RF+TFIDF	**0,7781 (0,0004)**	0,7589 (0,0026)

Tabela 7 – Resultados - F1-Score - Média (Variância)

SELEÇÃO - REPRESENTAÇÃO *versus* CLASSIFICADOR

Coleções de Referência (Representações: *Word2Vec* e *FastText*)				
Modelo	**R8**	**R52**	**RCV1**	**20 News**
SVM(RBF)+W2V	0,9239 (0,00030)	0,8486 (0,0006)	0,6451 (0,0002)	0,5065 (0,0377)
SVM(RBF)+W2V (E2V-IDF)	**0,9608 (0,00005)**	**0,9107 (0,0003)**	**0,7821 (0,0002)**	**0,6955 (0,0134)**
SVM(RBF)+FT	0,9188 (0,00030)	0,8433 (0,0007)	0,6360 (0,0003)	0,4957 (0,0313)
SVM(RBF)+FT (E2V-IDF)	**0,9616 (0,00004)**	**0,9091 (0,0002)**	0,7735 (0,0002)	**0,6705 (0,0140)**
KNN+W2V	**0,9599 (0,00010)**	0,9064 (0,0003)	**0,7905 (0,0002)**	0,6283 (0,0140)
KNN+W2V (E2V-IDF)	0,9562 (0,00010)	0,9064 (0,0004)	0,7746 (0,0003)	**0,6383 (0,0137)**
KNN+FT	0,9593 (0,00010)	0,9058 (0,0003)	**0,7859 (0,0003)**	0,6041 (0,0119)
KNN+FT (E2V-IDF)	0,9567 (0,00005)	**0,9070 (0,0003)**	0,7661 (0,0003)	0,6104 (0,0125)
DT+W2V	0,9106 (0,00020)	0,8079 (0,0006)	0,6351 (0,0002)	0,3948 (0,0072)
DT+W2V (E2V-IDF)	0,9152 (0,00030)	0,8098 (0,0008)	0,6241 (0,0002)	0,4156 (0,0068)
DT+FT	0,9154 (0,00010)	0,8037 (0,0006)	0,6292 (0,0002)	0,3489 (0,0053)
DT+FT (E2V-IDF)	0,9088 (0,00030)	0,8080 (0,0006)	0,6078 (0,0002)	0,3707 (0,0056)
RF+W2V	0,9522 (0,00020)	0,8767 (0,0005)	0,6985 (0,0002)	0,5156 (0,0136)
RF+W2V (E2V-IDF)	0,9467 (0,00010)	0,8787 (0,0005)	0,6870 (0,0002)	0,5337 (0,0126)
RF+FT	0,9498 (0,00010)	0,8770 (0,0005)	0,6904 (0,0002)	0,4743 (0,0114)
RF+FT (E2V-IDF)	0,9493 (0,00010)	0,8744 (0,0004)	0,6716 (0,0002)	0,4896 (0,0112)

Tabela 8 – Resultados - F1-Score - Média (Variância) - Treinados na Coleção

SELEÇÃO - REPRESENTAÇÃO *versus* CLASSIFICADOR

Coleções próprias (Representações: *Word2Vec* e *FastText*)	
Modelo	**Z5 News**
SVM(RBF)+W2V	0,7734 (0,0004)
SVM(RBF)+W2V (E2V-IDF)	**0,7838 (0,0003)**
SVM(RBF)+FT	**0,7738 (0,0006)**
SVM(RBF)+FT (E2V-IDF)	**0,7846 (0,0004)**
KNN+W2V	0,7648 (0,0002)
KNN+W2V (E2V-IDF)	0,7602 (0,0002)
KNN+FT	0,7642 (0,0002)
KNN+FT (E2V-IDF)	0,7620 (0,0002)
DT+W2V	0,6930 (0,0004)
DT+W2V (E2V-IDF)	0,6888 (0,0003)
DT+FT	0,6920 (0,0004)
DT+FT (E2V-IDF)	0,6770 (0,0003)
RF+W2V	0,7612 (0,0003)
RF+W2V (E2V-IDF)	0,7562 (0,0003)
RF+FT	0,7592 (0,0004)
RF+FT (E2V-IDF)	0,7518 (0,0004)

Tabela 9 – Resultados - F1-Score - Média (Variância) - Treinados na Coleção

SELEÇÃO - REPRESENTAÇÃO *versus* CLASSIFICADOR

Coleções de Referência (Representações: *Word2Vec* e *FastText*)				
Modelo	**R8**	**R52**	**RCV1**	**20 News**
SVM(RBF)+W2V	0,8359 (0,0014)	0,7731 (0,0010)	0,4972 (0,0003)	0,4831 (0,0323)
SVM(RBF)+W2V (E2V-IDF)	**0,9200 (0,0003)**	**0,8705 (0,0001)**	0,6937 (0,0002)	**0,5838 (0,0239)**
SVM(RBF)+FT	0,8361 (0,0015)	0,7674 (0,0011)	0,5019 (0,0002)	0,4822 (0,0331)
SVM(RBF)+FT (E2V-IDF)	**0,9279 (0,0003)**	**0,8759 (0,0001)**	0,6969 (0,0002)	**0,5859 (0,0256)**
KNN+W2V	0,9196 (0,0002)	0,8586 (0,0004)	**0,7416 (0,0003)**	0,5036 (0,0167)
KNN+W2V (E2V-IDF)	0,8716 (0,0004)	0,8133 (0,0003)	0,7004 (0,0004)	**0,5261 (0,0191)**
KNN+FT	**0,9251 (0,0002)**	**0,8626 (0,0004)**	**0,7470 (0,0003)**	0,5086 (0,0176)
KNN+FT (E2V-IDF)	0,8787 (0,0003)	0,8234 (0,0003)	**0,7070 (0,0005)**	0,5233 (0,0211)
DT+W2V	0,7948 (0,0011)	0,6844 (0,0012)	0,5414 (0,0001)	0,2539 (0,0046)
DT+W2V (E2V-IDF)	0,7365 (0,0007)	0,6167 (0,0017)	0,4985 (0,0002)	0,2579 (0,0060)
DT+FT	0,8047 (0,0006)	0,6882 (0,0013)	0,5457 (0,0001)	0,2684 (0,0049)
DT+FT (E2V-IDF)	0,7533 (0,0009)	0,6320 (0,0011)	0,5081 (0,0001)	0,2702 (0,0057)
RF+W2V	0,8710 (0,0004)	0,7747 (0,0009)	0,5903 (0,0001)	0,3498 (0,0120)
RF+W2V (E2V-IDF)	0,8249 (0,0007)	0,7284 (0,0011)	0,5427 (0,0002)	0,3625 (0,0145)
RF+FT	0,8769 (0,0006)	0,7848 (0,0009)	0,6011 (0,0001)	0,3683 (0,0140)
RF+FT (E2V-IDF)	0,8349 (0,0009)	0,7367 (0,0008)	0,5589 (0,0002)	0,3727 (0,0160)

Tabela 10 – Resultados - F1-Score - Média (Variância) - Pré-Treinado (Wiki)

SELEÇÃO - REPRESENTAÇÃO *versus* CLASSIFICADOR

Coleções próprias (Representações: *Word2Vec* e *FastText*)		
Modelo	**Z5 News**	**Z5 News Brasil**
SVM(RBF)+W2V	0,7553 (0,0006)	0,6679 (0,0010)
SVM(RBF)+W2V (E2V-IDF)	**0,7624 (0,0005)**	**0,7111 (0,0005)**
SVM(RBF)+FT	**0,7556 (0,0007)**	**0,7283 (0,0005)**
SVM(RBF)+FT (E2V-IDF)	**0,7641 (0,0005)**	**0,7501 (0,0008)**
KNN+W2V	0,7291 (0,0003)	0,5908 (0,0016)
KNN+W2V (E2V-IDF)	0,7055 (0,0002)	0,5986 (0,0017)
KNN+FT	0,7353 (0,0003)	0,6789 (0,0008)
KNN+FT (E2V-IDF)	0,7143 (0,0003)	0,6771 (0,0011)
DT+W2V	0,5836 (0,0005)	0,4151 (0,0016)
DT+W2V (E2V-IDF)	0,5550 (0,0003)	0,4093 (0,0015)
DT+FT	0,5996 (0,0004)	0,4609 (0,0018)
DT+FT (E2V-IDF)	0,5673 (0,0003)	0,4463 (0,0011)
RF+W2V	0,7037 (0,0004)	0,5537 (0,0014)
RF+W2V (E2V-IDF)	0,6771 (0,0006)	0,5520 (0,0017)
RF+FT	0,7107 (0,0004)	0,6063 (0,0015)
RF+FT (E2V-IDF)	0,6823 (0,0005)	0,5963 (0,0017)

Tabela 11 – Resultados - F1-Score - Média (Variância) - Pré-Treinado (Wiki + STIL)

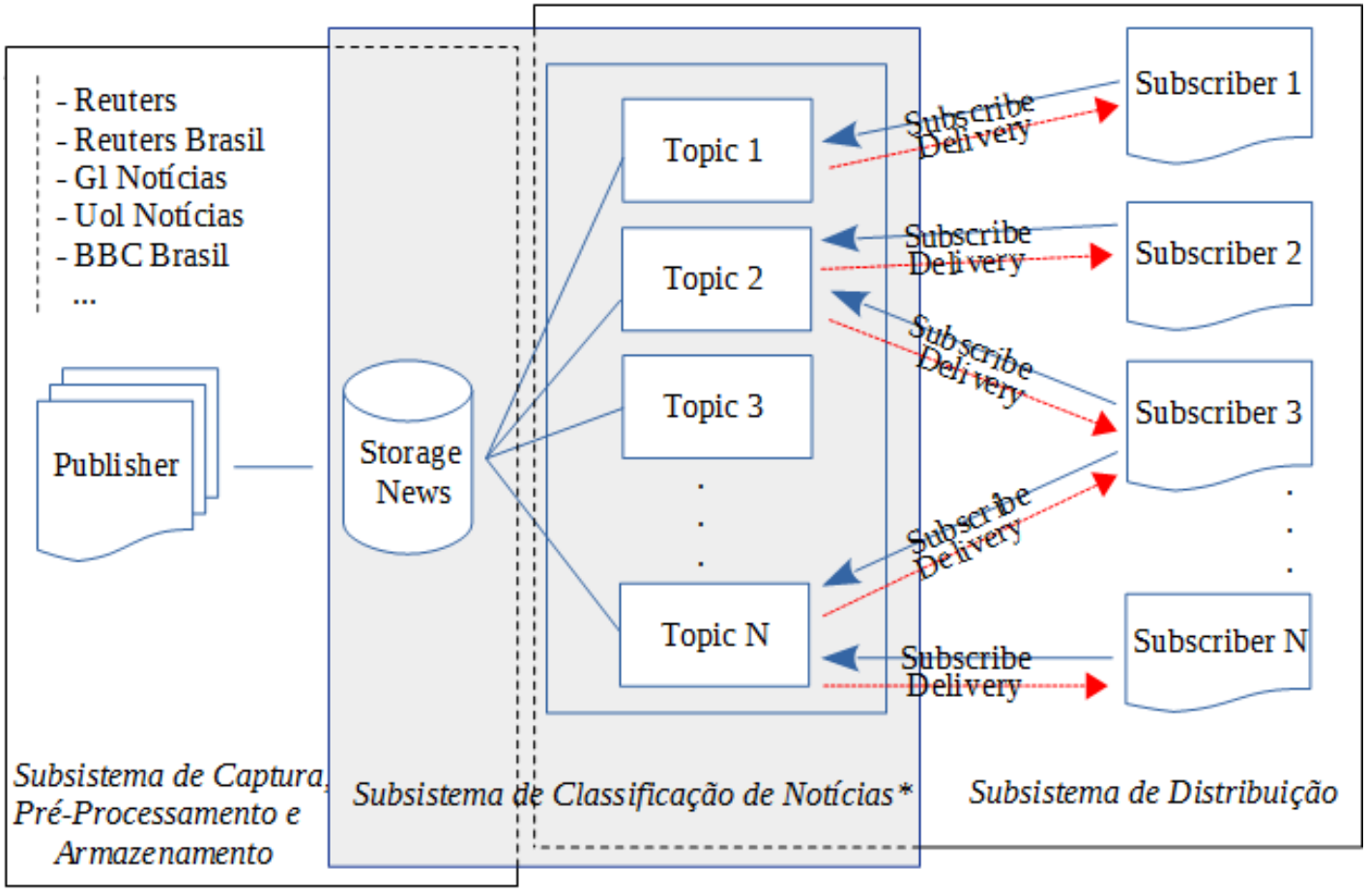
SELEÇÃO - REPRESENTAÇÃO *versus* CLASSIFICADOR

- BoW e TF-IDF - bem avaliadas RF (*Random Forest*) e SVM;
- BoW - Resultados melhores que *Embeddings* (Z5);
- *Embeddings* - Resultados melhores - Coleções Referências;
- *Embeddings* + `E2V-IDF` (**SVM**) - Melhor que *Média* (19%);
  - *Embeddings* Treinados na Coleção* x Pré-Treinados;
- Combinações mais bem avaliadas:
  - *SVM(RBF)+FT* (`E2V-IDF`);
  - *SVM(RBF)+W2V* (`E2V-IDF`);
  - *RF+BoW*.

Roteiro

LUPPARRECNEWS

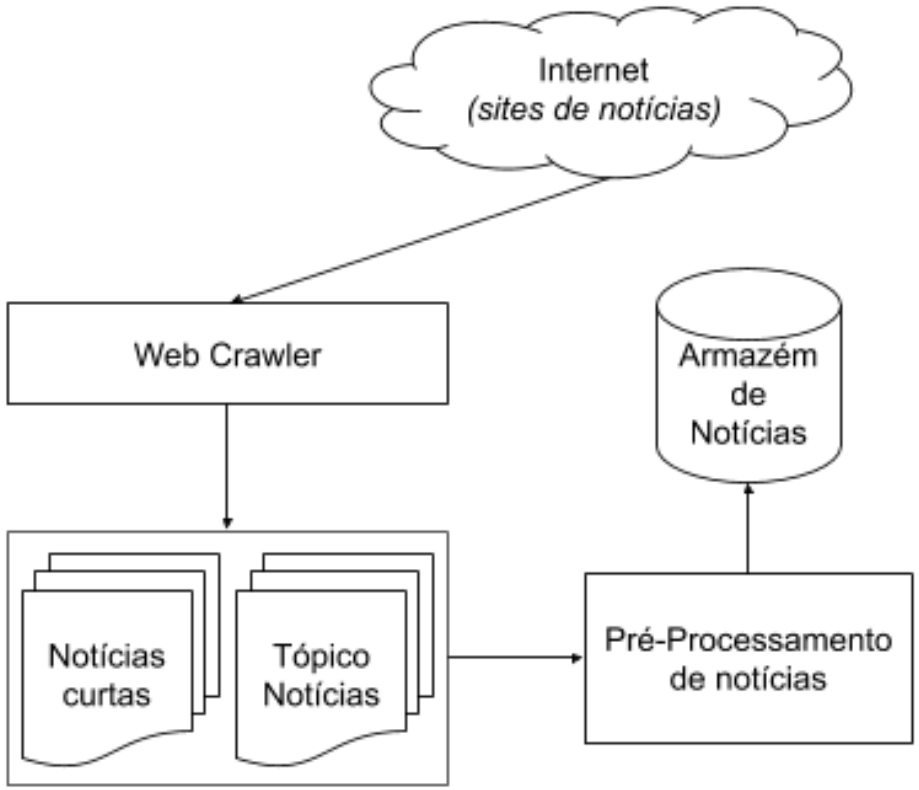
Estrutura do ***LupparRecNews*** baseada em *Publish-Subscribe*



Elaborado pelo autor

SUBSISTEMA DE CAPTURA, PRÉ-PROCESSAMENTO E ARMAZENAMENTO

Processo automático de coleta e transformação de notícias



Elaborado pelo autor

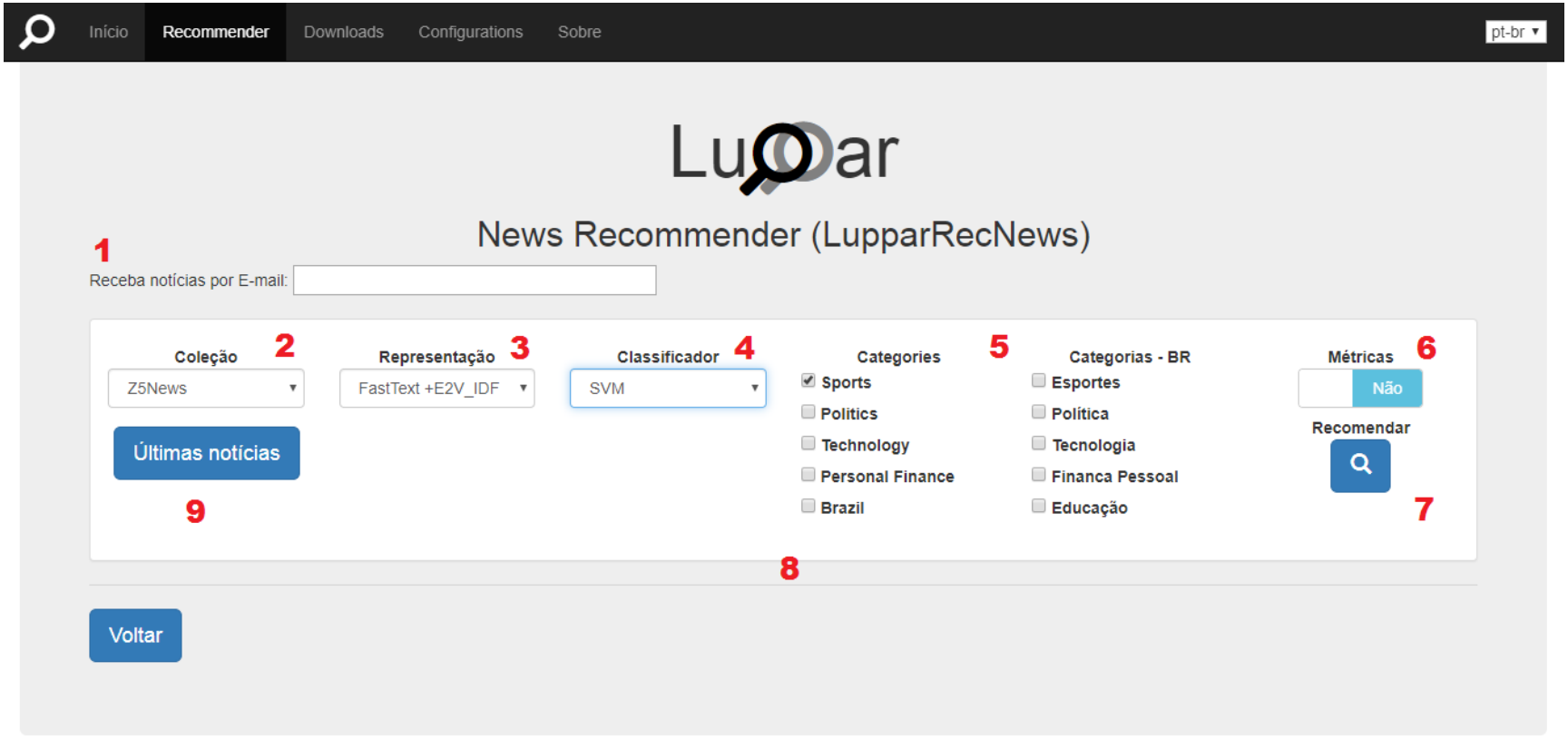
SUBSISTEMA: CLASSIFICAÇÃO / DISTRIBUIÇÃO

Subsistema de Classificação de Notícias (*core*) - Tem como objetivo classificar novos documentos (notícias) em tópicos (categorias) pré-estabelecidos.

Na fase de Avaliação e Seleção, as combinações mais bem avaliadas foram as seguintes respectivamente: *SVM(RBF)+FT* (`E2V-IDF`), *SVM(RBF)+W2V* (`E2V-IDF`) e *RF+BoW*. Implementadas no ***Luppar News-Rec***.

Subsistema de Distribuição - Abordagem para Recomendação: BASEADA EM CONTEÚDO (*subscrição em tópicos*) que tem como objetivo a distribuição dos documentos (notícias) para os usuários que assinaram os determinados tópicos.

LUPPARRECNEWS - INTERFACE



Em construção...

ROTEIRO

LUPPARRECNEWS - RESULTADOS

Z5News						
Modelo	**Precision**	**Recall**	**F1-Score**	**Acurácia**	**Retornados**	**RMSE**
RF+BoW	0,8710	0,5969	0,7083	0,5721	527	0,3135
SVM(RBF)+FT (E2V-IDF)	0,8349	0,6970	0,7597	0,6879	642	0,2839
SVM(RBF)+W2V (E2V-IDF)	0,8273	0,6853	0,7496	0,6762	637	0,2835

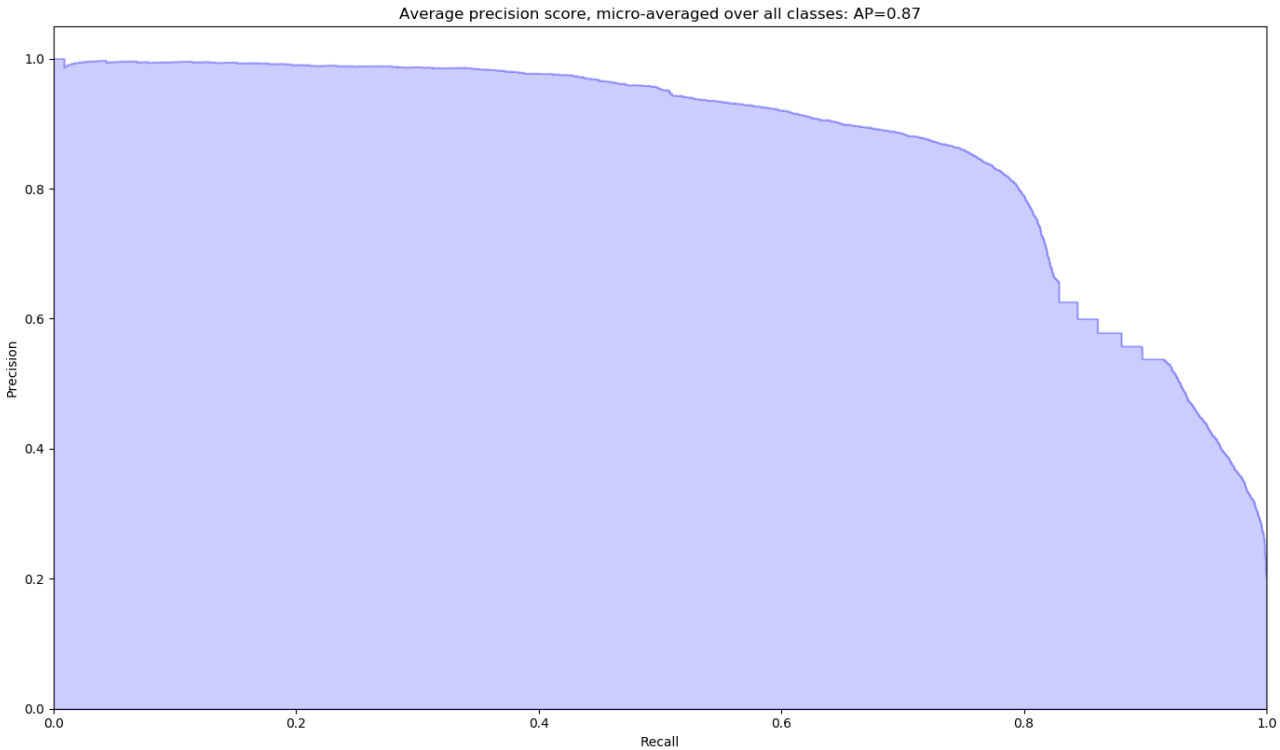
Tabela 14 – Comparativo de combinações - Tamanho do Teste: 769 notícias

Z5NewsBrasil						
Modelo	**Precision**	**Recall**	**F1-Score**	**Acurácia**	**Retornados**	**RMSE**
RF+BoW	0,8043	0,6326	0,7182	0,5976	516	0,3229
SVM(RBF)+FT (E2V-IDF)	0,8454	0,6250	0,7187	0,6250	485	0,3128
SVM(RBF)+W2V (E2V-IDF)	0,8030	0,5716	0,6679	0,5655	467	0,3372

Tabela 15 – Comparativo de combinações - Tamanho do Teste: 656 notícias

LUPPARRECNEWS - RESULTADOS (Z5NEWS)

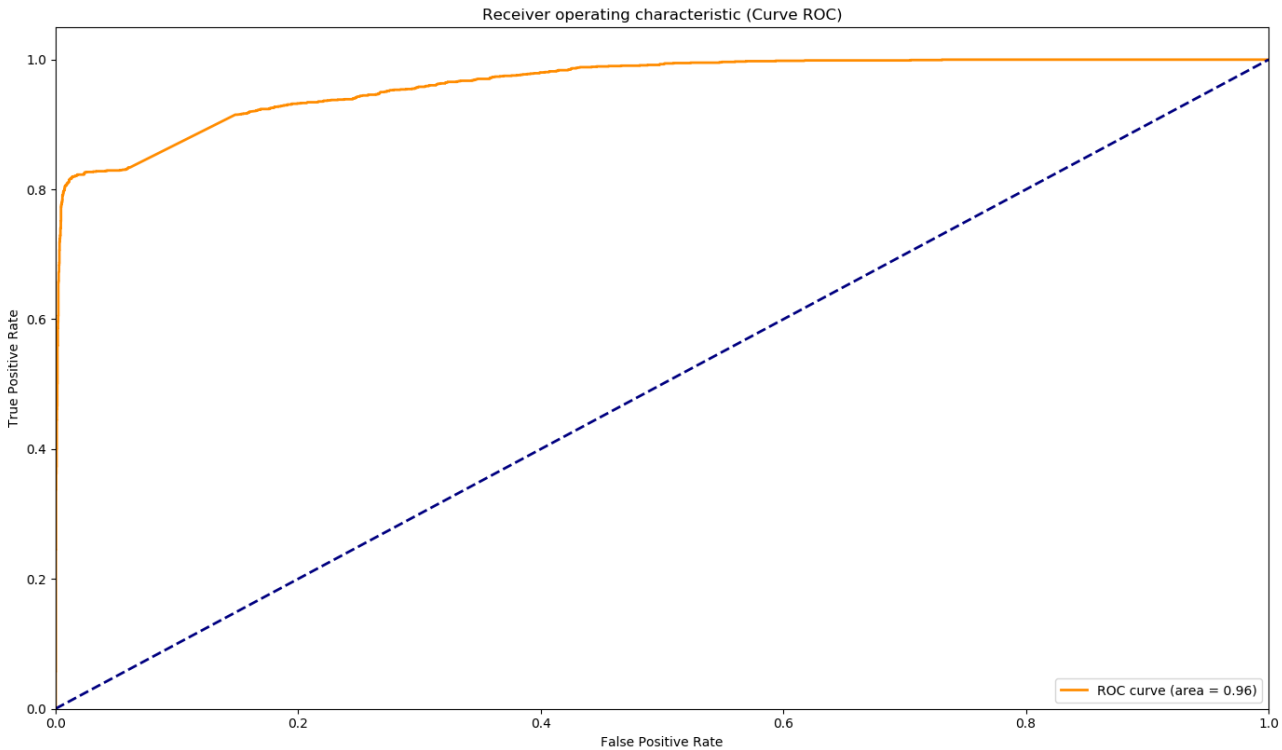
Área sob a Curva *Precision-Recall* (PR-AUC) - *Average Precision* (87%)



Elaborado pelo autor

LUPPARRECNEWS - RESULTADOS (Z5NEWS)

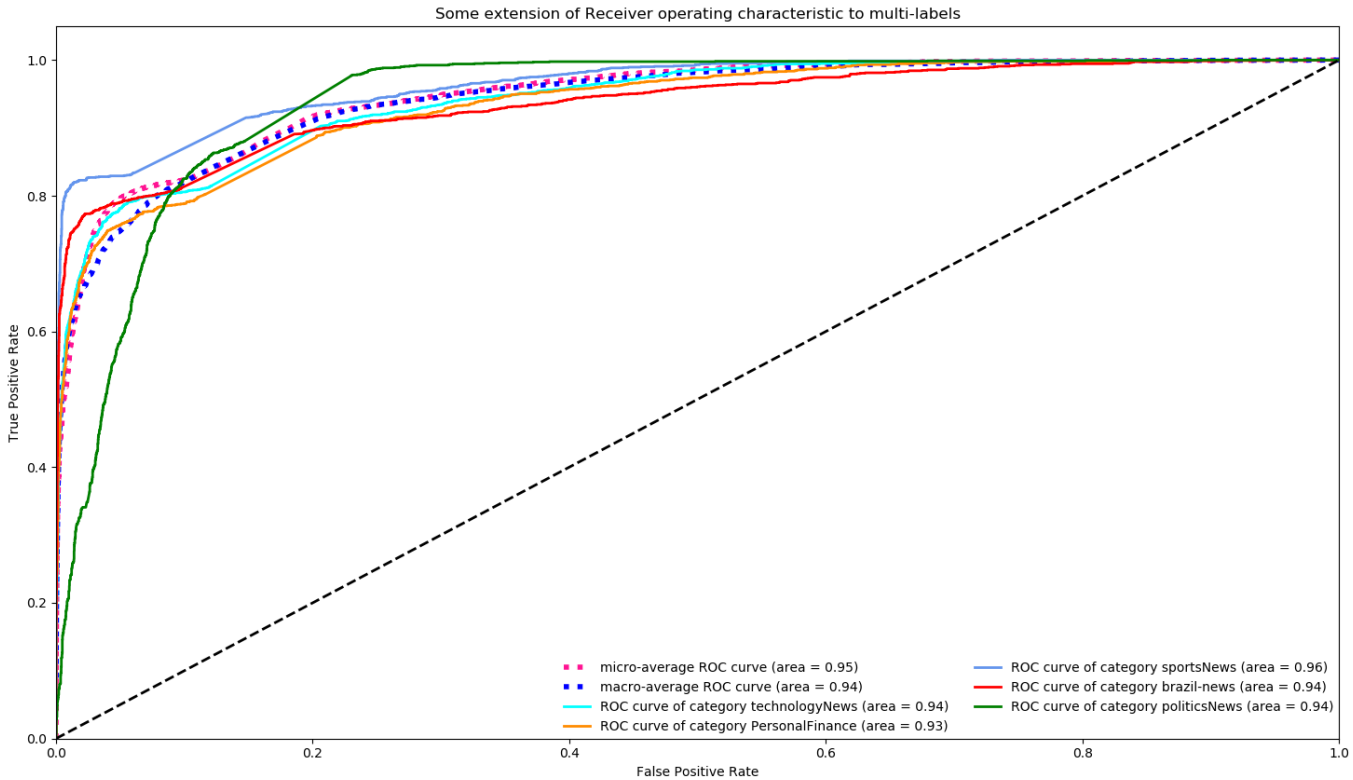
Curva ROC



Elaborado pelo autor

LUPPARRECNEWS - RESULTADOS (Z5NEWS)

Curva ROC detalhada por tópico



Elaborado pelo autor

ROTEIRO

LUPPARRECNEWS - DISCUSSÃO

As coleções de documentos podem ser categorizadas em:

- **Curtas** (Até 22 termos - Z5News e Z5NewsBrasil), obteve os melhores resultados combinando os classificadores com representações tradicionais: BoW e TF-IDF;
- **Médias** (Até 69 termos - R8 e R2), obtiveram os melhores resultados utilizando representações *Embeddings*;
- **Longas** (Até 141 termos - 20Newsgroups e RCV1), os melhores resultados também foram obtidos utilizando representações *Embeddings*.

LUPPARRECNEWS - DISCUSSÃO

Observe alguns exemplos de resultados (F1-Score) das tabelas 7 e 11, onde o fato de utilizar representações *Embeddings* na coleção de documentos própria Z5NewsBrasil (em Português) não superou os tradicionais BoW e TF-IDF, um ponto a se observar é que o vetor *embedding* foi pré-treinado e não treinado com base na coleção.

Já para a também coleção própria Z5News (em Inglês), que teve seu vetor *embedding* treinado com base na coleção, o melhor resultado foi utilizando a representação BoW, mas seguido de representações *Embeddings*, onde podemos também observar que a variância dos resultados dos *embeddings* são menores do que das representações tradicionais.

LUPPARRECNEWS - DISCUSSÃO

Note que ainda em relação a coleção própria Z5News, na tabela 12, é comparada as três combinações mais bem avaliadas, uma delas utilizando a representação tradicional BoW e as demais utilizando representações *Embeddings (Word2vec/FastText)*, vemos que os resultados para da métrica *Recall* foram maiores para as representações *Embeddings* e o número de documentos retornados (classificados) são maiores.

LUPPARRECNEWS - DISCUSSÃO

Essa combinação Classificador: SVM (*kernel*: RBF) e Representações *Embedding* ponderada por IDF (`E2V-IDF`) continua obtendo os melhores resultados na fase de testes do ***LupparRecNews***, como observamos nas tabelas 14, 15, 16 e 17.

Essas foram as melhores combinação obtidas neste trabalho, dentre essas combinações já definindo uma melhor, a combinação Classificador: SVM (*kernel*: RBF) e a representaçãos *Embedding: FastText* utilizando a abordagem ponderada por IDF (`E2V-IDF`) obteve os melhores resultados.

Roteiro

TRABALHOS FUTUROS

A evolução em andamento deste trabalho está sendo na implementação do Sistema Recomendador de Notícias com Aprendizagem de Perfil (Filtragem Colaborativa) e em melhorias na representação de documento `E2V-IDF`, para com isso tirar melhor proveito da densidade semântica de *embeddings* e obtenção de resultados ainda mais expressivos.

OBRIGADO