Universidade Federal de Juiz de Fora
Departamento de Ciência da Computação
Processamento de Linguagem Natural

Criando modelo de linguagem baseado em n-gram

Kleiton Ewerton de Oliveira

Professor: Jairo Francisco de Souza

Relatório do trabalho Criando modelo de linguagem baseado em n-gram

Juiz de Fora

Novembro de 2025

Sumário

1. Metodologia e Preparação dos Dados

Conforme estabelecido nos objetivos do trabalho, foi realizada uma prática de autoaprendizado focada na construção e análise de Modelos de Linguagem (LMs). Devido à indisponibilidade dos links originais sugeridos, optou-se pela construção de um *corpus* próprio extraído da Wikipedia em português.

- **Experimento completo**: Para acessar todos os scripts pythons usados, todos os comandos e todos os resultados acesse o repositorio `https://github.com/KleitonEwerton/language_models_with_SRILM_experiments_and_analysis`
- **Fonte de Dados:** Artigos aleatórios da Wikipedia (PT), processados via script Python.
- **Normalização:** O texto foi normalizado para *lowercase*, removendo pontuações excessivas e cabeçalhos Wiki, garantindo uma sentença por linha.
- **Separação dos Dados:**
  - **Treino:** 80% do *corpus* (utilizado para contagem de *n-grams*).
  - **Teste:** 20% do *corpus* (estritamente isolado para validação).
- **Vocabulário:** Vocabulário aberto derivado do treino.

```
kleitonoliveira@Tardis-Dell-G15:~/projetos/trabalho_lm$ head -n 10 treino.txt
martin sauer, secretário da expedição russa de 1791 que navegou às ordens de catarina
a turnê do álbum foi anunciada em 27 de outubro com início previsto para 2026.
no entanto, os efeitos foram mais graves no iêmen, onde as enchentes foram considerada
uma parte temporal seria algo como "o primeiro ano de vida de uma pessoa", ou "toda a
pacific philosophical quarterly, 62, p. 123–37.
demi também foi nomeada para a lista 30 under 30 da forbes na categoria música.
muitos historiadores afirmam que o último nome é mais adequado porque a vila de taruti
em 15 de setembro, demi anunciou o lançamento do seu nono álbum "it's not that deep" p
três pessoas morreram perto da cidade depois que as chuvas inundaram as ruas.
jung chung-rae, um deputado do partido democrático da coreia, criticou o nome por ser
```

Figura 1. Visualização das primeiras linhas do arquivo `treino.txt` demonstrando a normalização e tokenização.

2. Análise Intrínseca e Métricas Quantitativas

A primeira etapa da avaliação consistiu em medir a Perplexidade (PPL) e a robustez do modelo frente a dados desconhecidos.

2.1. O Problema da Esparsidade (Sparsity)

A análise dos dados brutos revelou uma alta esparsidade, típica de *corpora* pequenos e não curados manualmente.

- **Total de Palavras no Teste:** 7.087
- **Out of Vocabulary (OOVs):** 1.536 palavras (Aprox. 21,6%).
- **Singletons:** 5.795 *n-grams* apareceram apenas uma vez no treino.

```
$ awk '$2 == 1 {count++} END {print "Total de Singletons:", count}' treino.count
  Total de Singletons: 5795
```

Figura 2. Execução do comando `awk` para contagem de *singletons* no corpus de treino.

Análise: A alta taxa de *singletons* indicou precocemente que métodos sem suavização falhariam, pois a probabilidade de encontrar sequências inéditas no teste era estatisticamente certa.

2.2. Comparativo de Modelos e Suavização

Foram treinados modelos de ordem 3 (trigramas) variando as técnicas de desconto.

Tabela 1. Comparativo de Métricas por Modelo

Modelo	Método de Suavização	PPL	Falhas (ZeroProbs)	Observação
Baseline	Nenhuma (`-addsmooth 0`)	42.48*	1991	Valor enganoso. Modelo falhou em 28% das sentenças.
Laplace	Add-One (`-addsmooth 1`)	1143.12	0	Removeu falhas, mas degradou a precisão (PPL explodiu).
Kneser-Ney	Kneser-Ney Discount	246.86	0	Recuperação drástica da qualidade.
Witten-Bell	Witten-Bell Discount	239.33	0	Levemente superior ao KN neste dataset específico.
Híbrido	Interpolação Linear ($\lambda = 0.5$)	**232.09**	**0**	**Melhor Resultado.**

```
(venv) kleitonoliveira@Tardis-Dell-G15:~/projetos/trabalho_lm$ echo "--- BASELINE (Sem Suavização) ---"
--- BASELINE (Sem Suavização) ---
(venv) kleitonoliveira@Tardis-Dell-G15:~/projetos/trabalho_lm$ ngram -lm baseline.lm -ppl teste.txt
file teste.txt: 362 sentences, 7087 words, 1536 OOVs
1991 zeroprobs, logprob= -6385.8 ppl= 42.48151 ppl1= 62.19623
(venv) kleitonoliveira@Tardis-Dell-G15:~/projetos/trabalho_lm$ echo "--- LAPLACE (Add-One) ---"
--- LAPLACE (Add-One) ---
(venv) kleitonoliveira@Tardis-Dell-G15:~/projetos/trabalho_lm$ ngram -lm laplace.lm -ppl teste.txt
file teste.txt: 362 sentences, 7087 words, 1536 OOVs
0 zeroprobs, logprob= -18082.5 ppl= 1143.123 ppl1= 1809.347
(venv) kleitonoliveira@Tardis-Dell-G15:~/projetos/trabalho_lm$ echo "--- KNESER-NEY (O Melhor) ---"
--- KNESER-NEY (O Melhor) ---
(venv) kleitonoliveira@Tardis-Dell-G15:~/projetos/trabalho_lm$ ngram -lm kneser.lm -ppl teste.txt
file teste.txt: 362 sentences, 7087 words, 1536 OOVs
0 zeroprobs, logprob= -14146.63 ppl= 246.8665 ppl1= 353.5761
```

Figura 3. Saída do terminal mostrando os comandos de treino e os valores de PPL/ZeroProbs para os modelos Baseline, Laplace e Kneser-Ney.

```
(venv) kleitonoliveira@Tardis-Dell-G15:~/projetos/trabalho_lm$ ngram -lm baseline.lm -mix-lm unigram.lm -lambda 0.5 -ppl teste.txt
file teste.txt: 362 sentences, 7087 words, 1536 OOVs
0 zeroprobs, logprob= -13988.21 ppl= 232.0972 ppl1= 331.0881
```

Figura 4. Resultado do experimento de Interpolação Linear (Unigrama + Bigrama) demonstrando a redução da Perplexidade.

Discussão dos Resultados: O experimento confirmou que "números menores nem sempre são bons". O *Baseline* apresentou PPL 42.48, mas falhou em quase 2.000 contextos (*zeroprobs*), tornando-o inutilizável. A suavização de Laplace resolveu as falhas, mas introduziu ruído excessivo (PPL 1143). O equilíbrio ideal foi encontrado com métodos de desconto (Kneser-Ney/Witten-Bell) e Interpolação.

3. Avaliação Empírica (Qualitativa)

Para validar a utilidade real do modelo além dos números, foram realizados experimentos práticos de geração e classificação.

3.1. Visualização de Shannon (Geração de Texto)

Utilizando o modelo Kneser-Ney, frases foram geradas aleatoriamente baseadas nas probabilidades de *n-grams*.

> **Exemplo Gerado:** ”foi a cantora kehlani como shoyu”

```
(venv) kleitonoliveira@Tardis-Dell-G15:~/projetos/trabalho_lm$ python3 shannon_viz.py
Carregando kneser.lm...

--- Visualização de Shannon (Gerador de Frases) ---
Gerando 5 frases aleatórias:
1. lovato lançou seu apoio alex welch; sua carreira me levou a ser cozida e os moradores em uma colaboração entre
2. a peste de seul, bem acima da peste de 9 de gênero guerra mundial das nações unidas e 1490.
3. o declínio populacional tão importante liceu de setembro de emergência foi lançado em 2016, por um objeto y ao longo
4. ele entendia que causou a ser negada de junho em andhra pradexe.
5. demi deu a faixa na qual o seu divórcio de uma roupa estilo bondage deitada em quatro dessas cópias que

Gerando frases com início forçado ('o', 'a', 'foi'):
Início 'o': o terceiro álbum de crianças com o governo britânico.
Início 'a': a terceira posição da organização terrorista pelos eua, a campanha de outubro, a um deputado do sul do disco de 2020,
Início 'foi': foi lançado quando o martini cadet e seu primeiro dia 12 de pangim passou a epidemia de pangim passou a partir
```

Figura 5. Frases geradas aleatoriamente pelo script `shannon_viz.py`.

Análise: O modelo demonstrou coesão sintática local, mas falhou na semântica global. A conexão ”Kehlani”(cantora) com ”shoyu”(molho) é gramaticalmente possível, mas semanticamente improvável, evidenciando a ”memória curta”dos modelos de *n-grams*.

3.2. Teste de Acurácia (Frases Bem Formadas vs. Mal Formadas)

O modelo foi submetido a pares de frases para verificar se atribuía menor perplexidade (maior probabilidade) à frase correta.

- **Acurácia Global:** 75%
- **Caso de Sucesso:** ”o menino jogou a bola”(PPL 188) vs ”bola menino a jogou o”(PPL 599).
- **Caso de Falha:** ”a vida é bela”vs ”a vida é cadeira”(Empate técnico devido a OOVs).

```
(venv) kleitonoliveira@Tardis-Dell-G15:~/projetos/trabalho_lm$ python3 avaliacao_empirica.py
--- Avaliação Empírica: Acurácia do Modelo kneser.lm ---

Par 1:
  [Boa] 'o menino jogou a bola' -> PPL: 188.3339
  [Ruim] 'bola menino a jogou o' -> PPL: 599.7597
  Resultado: ACERTOU

Par 2:
  [Boa] 'a vida é bela' -> PPL: 135.0899
  [Ruim] 'a vida é cadeira' -> PPL: 135.0899
  Resultado: ERROU

Par 3:
  [Boa] 'hoje o dia está lindo' -> PPL: 380.1588
  [Ruim] 'lindo está dia o hoje' -> PPL: 384.086
  Resultado: ACERTOU

Par 4:
  [Boa] 'eu gosto de programar em python' -> PPL: 181.3954
  [Ruim] 'python de em gosto programar eu' -> PPL: 286.5502
  Resultado: ACERTOU

>>> Acurácia Final: 75.0%
```

Figura 6. Execução do teste de acurácia (`avaliacao_empirica.py`) exibindo os pares de frases e o diagnóstico de OOVs.

3.3. Classificador de Tópicos (Esporte vs. Música)

Foi implementado um classificador baseado na comparação de Perplexidade entre dois modelos distintos.

- **Resultado:** O classificador apresentou viés, classificando frases de Esporte como Música sistematicamente.

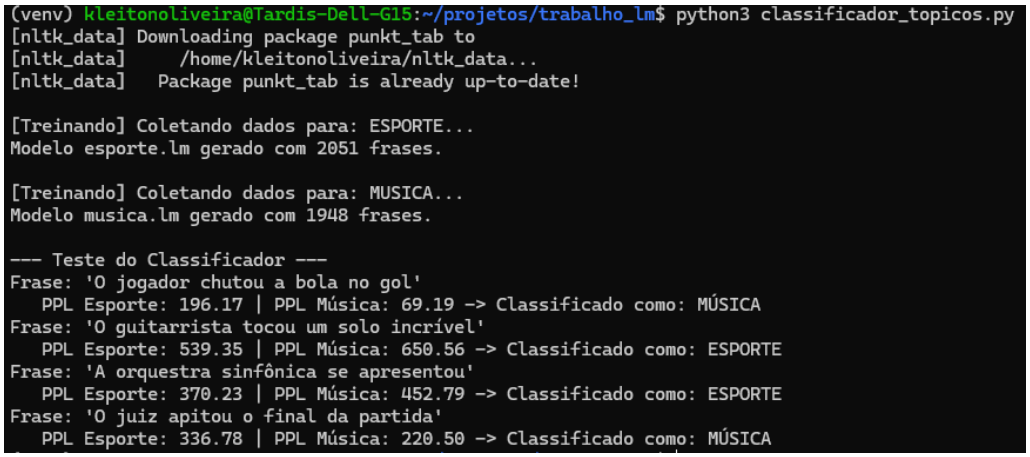

Figura 7. Resultado do classificador de tópicos (`classificador_topicos.py`) evidenciando o viés induzido pelo tamanho do vocabulário.

Análise Crítica: A falha não se deve à incapacidade dos *n-grams*, mas ao método de suavização combinado com vocabulários de tamanhos diferentes. O modelo de Música penalizou menos as palavras desconhecidas, gerando uma ”falsa confiança”(menor PPL) mesmo para frases de Esporte.

4. Conclusão

O trabalho permitiu constatar na prática as limitações da Estimativa de Máxima Verossimilhança (MLE) pura e a necessidade absoluta de Suavização (*Smoothing*) e *Backoff* para lidar com a esparsidade da linguagem natural.

Embora os modelos estatísticos (*n-gram*) sejam eficientes computacionalmente, os experimentos empíricos demonstraram suas fraquezas semânticas (Geração de Shannon) e dependência crítica da cobertura do vocabulário (OOV rate de 21%), sugerindo que para aplicações reais seria necessário um *corpus* de magnitude significativamente maior ou arquiteturas neurais.