Aproximação de textos artificiais utilizando Fontes de Informação de Markov

Eduardo Tenório
embat@cin.ufpe.br

Aproximação de textos artificiais utilizando Fontes de Informação de Markov

- Por que utilizar?
  - Fonte sem memória é restrita a poucas aplicações
  - Em problemas reais, o passado recente influencia o presente
  - Modelo probabilístico bem fundamentado
  - Fácil de implementar e executar

Aproximação de textos artificiais utilizando Fontes de Informação de Markov

- Onde é usada?
  - Estudo de linguagens naturais
  - Telégrafo
  - Aproximação de textos
  - Reconhecimento/Síntese de voz

Aproximação de textos artificiais utilizando Fontes de Informação de Markov

- Onde é usada?
  - Estudo de linguagens naturais
  - Telégrafo
  - Aproximação de textos
  - Reconhecimento/Síntese de voz

Aproximação de textos artificiais utilizando Fontes de Informação de Markov

- Onde é usada?
  - Estudo de linguagens naturais
    - Saber quais conjuntos de letras são mais utilizados
    - Entender a dependência entre fonemas e sílabas
    - Que letra é mais provável de aparecer após as *m* últimas?
    - “Antes de *P* e *B*, sempre usar *M*.”

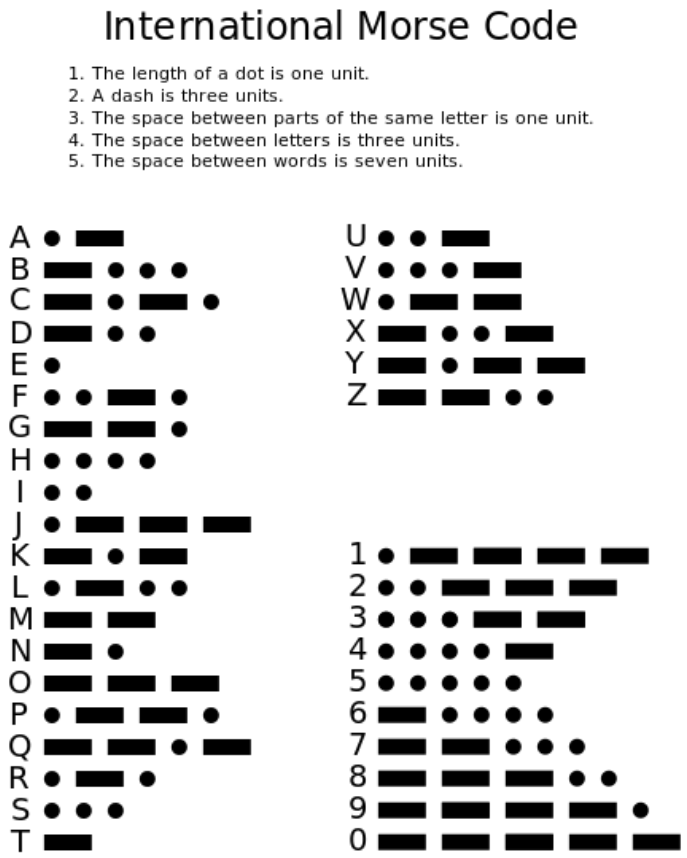
Aproximação de textos artificiais utilizando Fontes de Informação de Markov

- Onde é usada?
  - Estudo de linguagens naturais
  - Telégrafo
  - Aproximação de textos
  - Reconhecimento/Síntese de voz

Aproximação de textos artificiais utilizando Fontes de Informação de Markov

- Onde é usada?
  - Telégrafo
    - Códigos devem ser otimizados
    - Menos tempo e menos capacidade de canal requerida
    - Utiliza as estatísticas de um idioma (descritas no tópico anterior)
    - Código Morse
    - Código Baudot/Murray

Aproximação de textos artificiais utilizando Fontes de Informação de Markov



1 → e

3 → it

5 → san

7 → hurdm

9 → wgvlfbk

11 → opjxcz

13 → yq

Aproximação de textos artificiais utilizando Fontes de Informação de Markov

- Onde é usada?
  - Estudo de linguagens naturais
  - Telégrafo
  - Aproximação de textos
  - Reconhecimento/Síntese de voz

Aproximação de textos artificiais utilizando Fontes de Informação de Markov

- Onde é usada?
  - Aproximação de textos
    - Dado um idioma e uma fonte de Markov de ordem *m*, quanto mais relacionadas as letras e palavras, menos aleatório
    - Aumento do *m* → aumento da certeza
    - Diminui a entropia do sistema
    - Fica mais fácil decidir

Aproximação de textos artificiais utilizando Fontes de Informação de Markov

- Onde é usada?
  - Aproximação de textos (letras)
    - 0ª: 27 letras equiprováveis
      - “Markov” com apenas 1 estado
      - Todas as arestas são iguais
      - Apenas um amontoado de símbolos aleatórios
      - Não serve

XFOML_RXKHRJFFJUJ_ZLPWCFWKCYJ_FFJEYV
KCQSGHYD_QPAAMKBZAACIBZLHJQD

Aproximação de textos artificiais utilizando Fontes de Informação de Markov

- Onde é usada?
  - Aproximação de textos (letras)
    - 1ª: Letras com probabilidades distintas
      - “Markov” com apenas 1 estado
      - Arestas com valores diferentes
      - Ainda bem aleatório
      - Já é possível detectar algo fonético

OCRO_HLI_RGWR_NMIELWIS_EU_LL_NBNESEB
YA_TH_EEI_ALHENHTTPA_OOBTTVA_NAH_BRL

Aproximação de textos artificiais utilizando Fontes de Informação de Markov

- Onde é usada?
  - Aproximação de textos (letras)
    - 2ª: Digram
      - Markov de 1ª ordem
      - Um estado *i* possui uma probabilidade $p_i(j)$ de ir para o estado *j*
      - Probabilidade do digram: $p(i,j) = p(i)p_i(j)$

ON_IE_ANTSOUTINYS_ARE_T_INCTORE_ST_BE_S_DEAMY_ACHIN_D_ILONASIVE_TUCOOWE_AT_TEASONARE_FUSO_TIZIN_ANDY_TOBE_SEACE_CTISBE

Aproximação de textos artificiais utilizando Fontes de Informação de Markov

- Onde é usada?
  - Aproximação de textos (letras)
    - 3ª: Trigram
      - Markov de 2ª ordem
      - O estado *k* depende dos estados anteriores *i* e *j*.
      - $p(i,j,k) = p(i,j)p_{ij}(k) = p(i)\Sigma_j p_i(j)p_j(k)$
      - Mais que isso começa a ficar muito complexo

IN_NO_IST_LAT_WHEY_CRATICT_FROURE_BIRS
_GROCID_PONDENOME_OF_DEMONSTURES_O
F_THE_REPTAGIN_IS_REGOACTIONA_OF_CRE

Aproximação de textos artificiais utilizando Fontes de Informação de Markov

- Onde é usada?
  - Aproximação de textos (palavras)
    - 1ª Aprox.: REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE
    - 2ª Aprox.: THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED
    - Mais aproximado, mais complexo (número “infinito” de palavras). Podar?

Aproximação de textos artificiais utilizando Fontes de Informação de Markov

- Onde é usada?
  - Aproximação de textos
    - O idioma muda com o tempo
    - Tabela de frequências relativas de Robert Lewand (séc. 20) difere das frequências relativas extraída dos textos de Shakespeare (séc. 16)
    - Logo, basta atualizar a matriz de transição

Aproximação de textos artificiais utilizando Fontes de Informação de Markov

- Onde é usada?
  - Aproximação de textos (nível PhD)
    - Recentemente (Março de 2014) foi divulgado pelo MIT o uso de um sistema gerador de artigos científicos (SCIgen)
    - Um total de 120 artigos gerados por este sistema foram aceitos por periódicos (journals)
    - Puro *gibberish*

Aproximação de textos artificiais utilizando Fontes de Informação de Markov

- Onde é usada?
  - Estudo de linguagens naturais
  - Telégrafo
  - Aproximação de textos
  - Reconhecimento/Síntese de voz

Aproximação de textos artificiais utilizando Fontes de Informação de Markov

- Onde é usada?
  - Reconhecimento/Síntese de voz
    - Semelhante à fonte markoviana, mas com características adicionais
    - Utiliza Hidden Markov Model (HMM), modelo que segue a mesma idéia das fontes, porém com os estados intermediários desconhecidos (escondidos)
    - Bem mais complexo e um assunto para outra aula

Aproximação de textos artificiais utilizando Fontes de Informação de Markov

- Referências
  - Shannon, C. E. (1948), A Mathematical Theory of Communication. Bell System Technical Journal, 27: 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x
  - Norman Abramson (1963) Information Theory and Coding , New York: McGraw-Hill.
  - SCIgen: http://pdos.csail.mit.edu/scigen/

Aproximação de textos artificiais utilizando Fontes de Informação de Markov

- Referências
  - Reportagem sobre o SCIgen: http://www.natureworldnews.com/articles/6217/20140301/scholarly-journals-accepted-120-fake-research-papers-generated-by-computer-program.htm