Notes for *Speech and Language Processing*

Code2Hack

Contents

1 Introduction

2 Regular Expressions etc.

(P10)

2.1 Regular Expressions

2.1.1 Basic Regular Expression Patterns

- brackets: e.g [wW] means w or W. [a-z] means a to z. [ˆa-z] not a upper case letter. **Notice,** only when ˆ is the first character in brackets it negates the pattern.
- ?: /colou?r/ = color or colour. The preceding character or nothing.
- Kleene star: /[ab]*/= 0 or more a's or b's.
- Kleene +: /[0-9]+/= a sequence of digits.
- .: any single character. /.*/= any string
- Anchors:ˆ start of a line. $ = end. \b = boundary.

2.1.2 Disjunction, Grouping, Precedence

3 N-Gram Language Models

3.1 N-Grams

Here we denote w_1^{n-1} as the sequence $w_1, w_2, \dots, w_{n-1}$

$$P(w_1^n) = \prod_{k=1}^{n} P(w_k | w_1^{k-1})$$

For N-gram model which fufills N-1th-order Markov Property,

$$P(w_1^n) = \prod_{k=1}^{n} P(w_k | w_{k-N+1}^{k-1})$$