3 METODOLOGIA

Nos Capítulos anteriores, observa-se a importância dos AC, IA e RI para o projeto e desenvolvimento de um arcabouço de agente de conversação. Também foram descritas alguma ferramentas, tais como, AIML, modelo probabilístico de RI clássico e uma variação que seria o BM25, além da métrica MRR, utilizadas para projetar e desenvolver o AC proposto.

Este capítulo dedica-se a descrição da metodologia utilizada no Projeto e Desenvolvimento de um Arcabouço de Agente de Conversação aplicado ao curso de Sistemas de Informação da Pontifícia Universidade Católica de Minas Gerais.

O restante do capítulo encontra-se divido da seguinte maneira: a Seção 3.1 descreve a construção da base de documentos, na Seção 3.2 a modelagem do AC proposto é apresentada, a Seção 3.3 apresenta o interpretador utilizado para interpretação de marcações AIML; a Seção 3.4 descreve o MPC; a Seção 3.5 descreve o BM25; na Seção 3.6 é apresentado como o interpretador AIML faz comunicação com o *SWI-Prolog*; na Seção 3.7 é apresentada a comparação entre as três abordagens para busca de respostas; a Seção 3.8 descreve como o casamento e o tratamento de texto é feito; e, finalmente, a Seção 3.9 faz a descrição dos experimentos para validar o AC proposto utilizando a métrica MRR.

3.1 Construção da base de documentos

A construção de uma base de conhecimento grande é uma tarefa árdua se feita manualmente. Por isso, foi desenvolvido um programa na linguagem *C#* (*C Sharp*) para extração de informações da conta oficial da PUC Minas no sítio *http://ask.fm/pucminas*. Até o dia 23 de fevereiro de 2016 a conta da PUC Minas possuía 51820 perguntas/respostas. Neste sítio, pessoas (entre elas alunos, ex-alunos e vestibulandos) fazem perguntas que são respondidas por funcionários da universidade.

Pode se considerar que os documentos extraídos são relevantes e confiáveis, primeiro porque provêm de necessidades reais, segundo porque as respostas são escritas por profissionais capacitados.

As 51820 perguntas/respostas estavam armazenadas em 2073 páginas HTML. Estas páginas foram copiadas e salvas em documentos separados. Cada