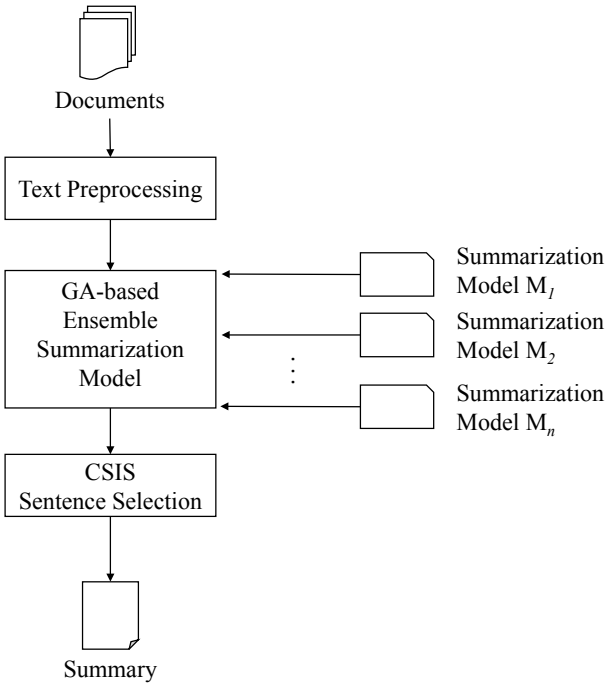

圖1、基因演算法組合摘要模型處理流程

要模型[15]，PageRank摘要模型[3]，以及iSpreadRank摘要模型[14]。它們的PLSA 主題強化摘要模型分別是PL-Degree，PL-NSDC，PL-PageRank 以及PL-iSpreadRank [13]。

（二）、文件前處理

在進行多文件摘要前，每份文件都需要先進行文本的前處理。由於文件敘述中會有一些缺乏意義的詞彙，例如冠詞或Be動詞等所謂的停用詞(Stopword)，這些詞通常無法表現文件摘要的語句重要性，因此在前處理中會被移除。以下是前處理中進行的項目：

- Sentence Segmentation：文件中的每一個語句先需要被斷句分開。本研究使用NLTK (Natural Language Toolkit)[1]工具來進行斷句。
- Tokenization：本研究使用NLTK以語句為單位，將語句中的單詞取出，以bag-of-words的方式來表示該語句。
- Stopword Removal: 我們使用Onix[2]停用詞表將語句中的停用詞移除。
- Stemming:在經過停用詞移除後，本研究再使用Porter Stemmer[3]進行詞根還原。

[1] https://www.nltk.org/
[2] http://www.lextek.com/manuals/onix/stopwords1.html
[3] https://tartarus.org/martin/PorterStemmer