

## Region models

영역(region) 모델이란 segments, extents, 또는 regions라 부르는 텍스트 데이터의 임의적 부분들을 추론(reason)하는 불리안 모델의 확장형이다. 이 모델은 a document collection에서 단어를 선형화하듯이 모델화 시킨다. 연속된 단어들의 특정한 순서를 영역이라 부른다. 영역들은 시작점과 끝점으로 구분한다. 다음의 예는 Shakespeare의 Hamlet의 일부분으로 단어위치에 번호가 부여되어 있다. 이 숫자는 단어 103에서 시작하여 단어 131에서 끝나는 영역을 보여주고 있다. 어구 “stand, and unfold yourself”는 이 텍스트에서 128번 위치에서 시작하여 131번 위치에서 끝나는 영역을 정의하고 있다. 어떤 영역들은 미리 정의할 수 있는데, 왜냐하면 이것들은 이 텍스트에서 논리적 요소를 표현하고 있기 때문이다. 예를 들어, Bernardo가 말한 글줄은 영역(122, 123)으로 정의된다.

<ACT>

<TITLE>ACT<sup>103</sup> I<sup>104</sup></TITLE>

<SCENE>

<TITLE>SCENE<sup>105</sup> I.<sup>106</sup> Elsinore.<sup>107</sup> A<sup>108</sup> platform<sup>109</sup> before<sup>110</sup> the<sup>111</sup> castle.<sup>112</sup></TITLE>

<STGDIR>FRANCISCO<sup>113</sup> at<sup>114</sup> his<sup>115</sup> post.<sup>116</sup> Enter<sup>117</sup> to<sup>118</sup> him<sup>119</sup> BERNARDO<sup>120</sup></STGDIR>

<SPEECH>

<SPEAKER>BERNARDO<sup>121</sup> </SPEAKER>

<LINE>Who's<sup>122</sup> there?<sup>123</sup></LINE>

</SPEECH>

<SPEECH>

<SPEAKER>FRANCISCO<sup>124</sup> </SPEAKER>

<LINE>Nay,<sup>125</sup> answer<sup>126</sup> me:<sup>127</sup> stand,<sup>128</sup> and<sup>129</sup> unfold<sup>130</sup> yourself.<sup>131</sup></LINE>

영역 시스템은 다큐검색에만 제한되진 않는다. 어플에 따라, 우리는 텍스트 쿼리를 사용하여 각본 전체를 탐색하고자 할 수도 있고, 말하는 사람과 관련된 장면을 탐색할 수도 있고, 말하는 사람의 말 내용을 검색하고자 할 수도 있으며, 인용부호를 사용하거나 말하는 사람과 관련된 단일 글줄을 탐색하고자 할 수도 있다. 불리안 모델은 다큐가 명사형 식별자(nominal identifier)로 표현되는 다큐의 집합에서 운영된다는 것을 생각해 보면, 영역모델은 영역이 두 개의 서수 식별자((ordinal identifiers): 다큐 콜렉션에서 시작점과 끝점)로 표현되는 영역의 집합에서 운영되는 모델로 생각할 수 있다. 불리안 연산자 AND, OR, NOT과 같은 교집합, 합집합, 차집합으로 영역의 집합을 직접 정의할 수 있다. 영역모델은 적어도 2개 이상의 연산자를 사용한다: CONTAINING과 CONTAINED BY. 영역 쿼리를 지원하는 시스템은 Hamlet이 “farewell”이라고 말한 모든 글줄을 검색하도록 하는 다음과 같은 복잡한 쿼리를 처리할 수 있다.

(<LINE> CONTAINING farewell) CONTAINED BY (<SPEECH> CONTAINING  
<SPEAKER> CONTAINING Hamlet))